DOCUMENT RESUME

ED 266 156 TM 860 099

AUTHOR	Burstein, Leigh; And Others
TITLE	Using State Test Data for National Indicators of Education Quality: A Feasibility Study. Final Report.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	Nov 85
GRANT	NIE-G-83-001
NOTE	275p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC11 Plus Postage.
DESCRIPTORS	Data Collection; *Educational Assessment; *Educational Quality; Elementary Secondary Education; Feasibility Studies; Longitudinal Studies; Mastery Tests; Minimum Competency Testing; National Norms; Pilot Projects; Public Policy; Standardized Tests; *State Programs; *State Surveys; *Testing Programs
IDENTIFIERS	*Educational Indicators; National Assessment of Educational Progress

ABSTRACT

The desire for a national picture of educational quality remains a continuing but unresolved goal. A question has been raised among high level policymakers regarding the feasibility of using existing data collected by the states to construct education indicators for state-by-state comparisons of student performance at the national level. A feasibility study was contracted to the UCLA Center for the Study of Evaluation (CSE) to explore the methodological and implementation issues of this approach. The results of the feasibility study are described and discussed in this report. Included in the study are analyses of: (1) the general characteristics of current state testing programs and of the content of currently used state tests; (2) alternative approaches to linking test results across states to create a common scale for purposes of comparison; and (3) the availability of auxiliary information about students and schools and its potential use in creating more valid indicators of achievement. These analyses culminated in a number of recommendations about ways to facilitate the use of state data for national comparisons. These recommendations focus on basic preconditions, proposed approaches, pilot study needs, auxiliary information collection and documentation, and strategies for optimizing political, institutional, and economic support. (Author/LMO)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Center for the Study of Evaluation

UCLA Graduate School of Education
Los Angeles, California 90024

USING STATE TEST DATA FOR NATIONAL INDICATORS OF EDUCATION QUALITY: A FEASIBILITY STUDY

Leigh Burstein, Eva L. Baker,
and Pamela Aschbacher
Center for the Study of Evaluation
University of California, Los Angeles

J. Ward Keesling
Advanced Technology, Inc.

ED266156

TM 860 099

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
C. Griffith

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

FINAL REPORT

USING STATE TEST DATA FOR NATIONAL INDICATORS OF EDUCATION QUALITY: A FEASIBILITY STUDY

Leigh Burstein, Eva L. Baker,
and Pamela Aschbacher
Center for the Study of Evaluation
University of California, Los Angeles

J. Ward Keesling
Advanced Technology, Inc.

November, 1985

Grant Number NIE-G-83-001

Center for the Study of Evaluation,
Graduate School of Education,
University of California at Los Angeles

ACKNOWLEDGEMENT

This study was funded by Grant NIE-G-83-001 from the National Institute of Education and jointly monitored by the National Center for Education Statistics. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education or the National Center for Education Statistics, and no official endorsement by either agency should be inferred.

We wish to thank Emerson J. Elliott for his initiation and support for the project. Conrad Katzenmeyer served as NIE project monitor with Jean Brandes of the NCES assistance in facilitating interagency communication. The Policy and Technical Panel members (Darrell Bock, Dale Carlson, Ward Keesling, Tom Kerins, Robert Linn, E. Roeber, Richard Shavelson, Lorrie Shepard and Marshall Smith) contributed substantially to this report through their advice on study options, written comments,and support for a thorough and open exploration of complex technical and political issues. Bill Doherty supervised the collection of the Telephone Interview Survey data from State Testing Programs. State Testing Directors were very cooperative in providing information and copies of the reports they distributed. Victoria Gouveia, Kathy Fliller, and Melinda Baccas provided secretarial, administrative, and research assistance, and Ms. Gouveia was responsible for graphical layout and report assembly. Fiscal management and administration was ably provided by Dr. Joan Herman of CSE.

TABLE OF CONTENTS

Executive Summary

Chapter 1. Project Overview

Purpose of Study 1.1
Project Activities 1.2
Recommendations from First Policy and Technical Panel Meeting 1.3
Recommendations of Second Policy and Technical Panel Meeting 1.4
Overview of the Report 1.6

Chapter 2. Description of Existing State Testing Programs

Procedures 2.1
    State Participation 2.1
    Focus of Interview 2.2
Summary of State Testing Activities 2.4

Chapter 3. Consideration of Common Test Linking Strategies

Statement of the Problem 3.1
Procedures for Examining Alternative Approaches 3.2
Basic Psychometric Alternatives 3.3
    Matched Test Data 3.4
    Common Anchor Items 3.5
    Preferred Option 3.7
Source of Common Anchor Items 3.7
    NAEP 3.8
    Commercially Available Standardized Tests 3.9
    State Developed Items 3.10
    Other Sources 3.11
    Preferred Option 3.12
Implementation Issues 3.13
Summary and Recommendations 3.15

Chapter 4. Content Analysis of Existing State Tests

Statement of Problem 4.1
Procedures 4.3
Basic Results 4.7
    Reading 4.8
    Mathematics 4.13
    Writing 4.18
Exemplary Practices 4.18
Summary and Recommendations 4.23

Chapter 5. Examination of Reporting Practices and Auxiliary Information

Statement of the Problem 5.1
    Longitudinal Contrasts 5.1
    Subgroup Contrasts 5.2
Current Collection and Reporting Practices 5.2
Summary and Recommendations 5.7

Chapter 6. Overall Summary and Recommendation

Preconditions and Guiding Principles 6.2
Pilot Study 6.4
Auxiliary Information and Documentation 6.5
Political, Institutional, and Economic Environment 6.6
Cost Implications: An Addendum 6.7

Appendices

1. List of Panelists for Feasibility Study
2. Telephone Interview Guide
3. Partial Summary of First Policy and Technical Panel Meeting
4. Decision Memorandum on the Feasibility of Using State-level Data for National Educational Quality Indicators
5. Sources of Information about State Testing Programs
6. Survey Summary of General Characteristics of State Testing Programs
7. Bock Memorandum and Panel Responses on Common Test Linking Issues
8. Summary of Documents Provided by State Testing Programs
9. Master Matrices for Math, Reading, and Writing
10. Decision Rules for Content Analysis of State Tests
11. Rating Categories of "Source Quality"
12. Comments on Sources of Information and Quality of Information
13 Definition and Identification of Skills
14. Key to Summary Sheets
15. Detailed Reading Summary from Content Analysis of State Tests
16. Detailed Math Summary from Content Analysis of State Tests
17. Detailed Writing Summary from Content Analysis of State Tests
18. Summary of Number of Items and Subskills in Each Cell of the Math Matrix for Grades 4-6 and 4-9 in AL, CA, FL, LA, and PA
19. Coding of Reporting Practices and Auxiliary Information
20. Linking State Educational Assessment Results: A Feasibility Trial

STATE TESTS AS QUALITY INDICATORS PROJECT

EXECUTIVE SUMMARY

The desire for a national picture of educational quality remains a continuing but unresolved goal. Last fall, a question was raised among high level policymakers regarding the feasibility of using existing data collected by the States to construct education indicators for state-by-state comparisons of student performance at the national level. A feasibility study was contracted to the UCLA Center for the Study of Evaluation (CSE) to explore the methodological and implementation issues of such an approach.

The results of the feasibility study are described and discussed in this report. Included in the study were analyses of the general characteristics of current state testing programs and of the content of currently used state tests; of alternative approaches to linking test results across states to create a common scale for purposes of comparison; and of the availability of auxiliary information about students and schools and its potential use in creating more valid indicators of achievement.

These analyses culminated in a number of recommendations about ways to facilitate the use of state data for national comparisons. These recommendations focus on basic preconditions, proposed approaches, pilot study needs, auxiliary information collection and documentation, and strategies for optimizing political, institutional, and economic support.

The following recommendations are made regarding basic preconditions and guiding principles for the use of state test data:

1. The comparison of the performance of states should include only those states where there is sufficient empirical evidence to allow analytical adjustments for the effects of differences in testing conditions. All states that collect test data on the pertinent content areas at the designated grade levels or whose test results can be statistically adjusted to the targeted testing conditions should be considered for inclusion in cross-state comparisons.

2. Existing state testing procedures should be disrupted as minimally as possible. Only those data collection activities considered essential for obtaining evidence of comparability should be introduced over and above the states' own planned expansions and extensions of their testing activities.

3. Existing state tests and testing data should be used as much as possible.

4. Regardless of the optimal specificity desired in the reporting of cross-state performance, the content of the tests to be used for comparison purposes should be specified at as low a level (subskill or subdomain) as possible to enhance the quality of the match to existing tests and to encourage attention to the content and detail of what is being tested.

5. If cross-state comparisons are to be achieved through linking of a state's test to a common linking test, the content covered by the linking test should be as broad as possible both to ensure overlap with each state's tests and to encourage broadening rather than narrowing of the curriculum across the states.

6. The proposed approaches for developing state-by-state achievement indicators should be compatible with the wider issue of the development of systems for monitoring instruction practices as well as educational progress both within and across the states. Desireable augmentations of current state practices should increase documentation of student and school characteristics within the framework of planned changes in state educational activities.

The following recommendations are made with regard to optimal approaches to the problem of linking test data across states and the implementation of the desired approaches.

7. A common anchor item strategy, wherein a common set of linking test items is administered concurrently with the existing state test to an "equating-size" sample of schools and students, should be used as the basis for expressing test scores from different states on a common scale.

8. The items contributing to the common anchor set should be selected from multiple sources including existing state-developed tests, NAEP, commercially available tests, and other policy relevant and technically adequate sources, such as the IEA tests.

9. The mechanisms for establishing the skills to be included in the common anchor set, for selecting items to represent the skills, and for specifying the rules for participation by individual states should be developed and administered primarily by collective representation of the states.

10. The organization responsible for developing and administering the linking effort should consider the following points relevant to implementation:

a. Procedures for documenting contents of existing state tests should be specified so that questions of what is being equated to what can be addressed.

b. Specification of content represented in common anchor set should be at the lowest level possible (subskill level) even if achievement indicators, at least initially, are to be reported at higher levels (skill or content area).

c. The minimum criteria for considering an item for inclusion in the common anchor item set should be that

- o The item measures a skill selected for inclusion in the common anchor item set, and
- o Sufficient empirical evidence is available about the item to ascertain its behavior for the major segments of the student population with which it will be used.

d. The selection of items should be made by teams of curriculum and testing specialists from a broad-based pool of items without identification of their source as is technically feasible.

e. The following set of testing conditions should be specified:

- o Target grades and range of testing dates along with requirements for special studies in those states who normally test outside the chosen range or do not test at present but elect to participate.
- o Procedures for concurrent administration of the common anchor item set with existing state tests for the various alternative types of state tests (matrix sampled,state-developed single form, commercially developed standardized test).
- o Auxiliary information for checking subgroup bias and determining sample representativeness (for equating and scaling purposes).
- o Minimum sample sizes (for both schools and students).

The following recommendation is made with regard to the need for pilot studies of the proposed approach:

11. A pilot study of the proposed common test linking strategy should be conducted in a limited set of skill areas for a specific grade range in order to determine both the quality of the equating under preferred conditions and the effects of various deviations from these conditions. The content areas and grade levels to be used in the proposed pilot study are literal comprehension for reading and either numbers and numeration or measurement for mathematics at grades 7-9.

The following recommendations are made with regard to the need for auxiliary information and documentation about student and school characteristics:

12. The organization responsible for coordinating the test linking activities described earlier should also develop plans for obtaining routinely a select set of common auxiliary information from states about their students and schools.

13. Cooperating states should be encouraged to provide on an annual basis uniform documentation describing their data collection activities.

14. Cooperating states should work toward the collection of a common set of auxiliary information about student and school characteristics along with their testing data. A standard set of definitions for measuring the chosen characteristics should be determined.

15. The organization responsible for coordinating test linking efforts should consider ways of contextualizing state test comparison data to mitigate against the possibility of unwarranted interpretations. The auxiliary information gathered as part of the previous recommendation should contribute to this activity.

The following recommendations are made with regard to establishing an effective political, institutional, and economic environment for the indicator effort:

16. To develop the necessary levels of political support for this activity, broad-based support for the idea should be developed. Key participants include Chief State School Officers, their staffs,and other state education officials; other prominent state officials, including the Governor, Members of Congress, and state legislators; and representation of members of large city school districts, the education associations and from the private sector.

17. An institutional structure for the conduct of this activity that relies heavily on the collective efforts of the states should be adopted. The Council of Chief State School Officers' new Assessment and Evaluation Coordinating Center proposal deserves consideration for this purpose.

18. Technical assistance and oversight should be established to assure the technical and methodological quality of the linking and equating, of the content of measures, and of validity of interpretations. This oversight should be provided by independent or semi-independent panels, perhaps modeled on the panels advising the NAEP activity.

19. A long-term, secure basis of financial support for coordinating and updating the test linking activity and the collection and reporting of common auxiliary information should be developed. This support is necessary to ensure that modifications in the basis of comparison and in the participating states can be accommodated over time while maintaining the integrity of the linking effort.

Chapter 1
Project Overview

Purpose of Study

Various efforts to improve the capacity for collecting and reporting achievement indicators of educational quality and to improve methods for obtaining comparable state-level performance data serve as both a backdrop and an impetus for this study. One natural consequence of both the recent concern for the quality of existing educational offerings and the desire to monitor the consequences of proposed reforms has been an expanded search for high quality data to inform educators and policy makers. Various groups have begun to search for education indicators to serve as benchmarks for judging educational progress and status. Former Secretary of Education Bell's release of his State Education Statistics charts with state data and state rankings on the SAT and ACT plus other variables is the most visible example of this effort. The attention it received from the press, the public, and various education organizations established the current climate in which other education indicator efforts are viewed.

Of particular concern in the realm of indicators of educational performance has been the appropriate selection and proper use of measures of educational achievement to compare the accomplishments of individual states. A basic dilemma is that although students undergo a substantial amount of testing during the course of their educational careers, virtually all of this testing is determined by local and state policies (annual district standardized achievement testing, state assessments, minimum competency and proficiency testing) or by individual need and initiative (special education testing, college admissions examinations). While these testing activities may be suitable for the purposes for which they were designed, none can be readily translated into a uniformly acceptable achievement standard for comparing the quality of educational programs across states. In essence there exists no nationally common test that is currently administered in a manner that will serve such a purpose. The self-selection in taking the SAT and ACT make their results a flawed basis for state-level comparisons. The current design for sample selection and administration schedule of the National Assessment of Educational Progress (NAEP) does not provide sufficiently representative or current data in most states to make it a suitable source for such comparisons.

The desire for a national picture of educational quality remains a continuing but unresolved goal. In the past, there has been some resistance from States about comparative information of any sort. The arguments have centered on the need for good contextualization of information so that differences in performance can be properly attributable to quality of educational services and not to social and economic conditions in the regions themselves.

A national test has been proposed periodically as a solution, but has been rejected because of the constitutional delegation of educational responsibilities to the States and the attendant notion that such a test would exert untoward Federal pressures toward uniformity in educational practices. The cost of such a new test (or radical expansion of the NAEP sampling and scheduling) would also be high.

Last fall, a question was raised among high level policymakers regarding the feasibility of using existing mechanisms within the States to contribute to the picture of American educational quality. Specifically under consideration was the extent to which existing measures of student performance collected by the States could be combined to 1) provide a national profile of performance in achievement domains; 2) provide a basis for state-by-state comparisons of student performance. A feasibility study (hereafter referred to as the State Tests as Quality Indicators (STQI) Project) was contracted to the UCLA Center for the Study of Evaluation (CSE) to explore the methodological and implementation issues of such an approach. This report describes the activities of the STQI Project, summarizes project analyses, and presents recommendations regarding the feasibility of using existing state tests for the desired purposes.

Project Activities

The basic charge to CSE in conducting the STQI Project was to document existing state testing program activities with specific emphasis on the possibilitity of using data already routinely collected to form "comparable" state-level achievement indicators and to determine the analytical and psychometric methods necessary or potentially appropriate to generate the desired indicators. With respect to the latter, the original proposal identified four general approaches that might be applicable: direct equating of test content; econometric adjustments for selection and/or economic and socioeconomic conditions; equating by the use of a common test or linking measure; and methods that depend only on within-state information such as trend data and subgroup comparisons.

To implement its charge, CSE carried out the following activities:

1. Conducted a telephone interview survey of State testing directors to obtain information about their program characteristics;

2. Examined copies of reports routinely generated by the State testing programs to ascertain additional details about the content being assessed and the procedures used for analyzing and reporting results;

3. Convened two panel meetings of scholars and practitioners in Washington (November 29-30, 1984; April 15-16, 1985) to engage

in a discussion of issues and options along with interested observers from government and professional organizations.

4. In response to a modified charge coming out of the first Panel meeting, carried out a detailed content analysis of existing state tests (both state-developed and commercially developed) and

5. Identified the nature and range of auxiliary information about student and school characteristics either collected or reported with state testing data that might serve as additional factors to consider with respect to the quality of a state's educational performance.

The details of activities 1,2,4, and 5 are reported in subsequent chapters. To provide perspective on the reasons for these activities, it is necessary to recount the recommendations coming out of the two panel meetings and CSE actions in response to the recommendations.

Recommendations from the First Policy and Technical Meeting

The Policy and Technical Panel for the STQI Project (A complete list of panelists is provided in Appendix 1) included university scholars with both policy and technical expertise relevant to the project's focus and practitioner representatives from several major long-term state testing programs. The meetings of the Panel were scheduled in Washington so that representatives of the governmental agencies with interest in education indicators (National Center for Education Statistics, National Institute of Education, Office of Planning, Budget, and Evaluation, Office of Technology Assessment) and various professional organizations could participate in the discussions.

The purpose of the first Panel meeting was to consider which of the available approaches for deriving indicators from state data were potentially useful given current testing practices, and thus which approaches CSE should explore in greater depth using reports provided by the states. As preparation for the meeting, CSE conducted in-depth telephone interviews (Appendix 2) with representatives from state testing programs and requested copies of existing reports and content specifications generated by state testing programs. The results of these phone interviews were then combined with information from other recent surveys of state testing activities and distributed to meeting participants. This information was intended to place the proposed approaches within a context of existing practices and aid in the effort to refine and focus the remaining tasks of the feasibility study.

A partial summary of the deliberations at the first Panel meeting is provided in Appendix 3. While there was interest in all the approaches considered for combining state-level data for national comparative purposes, opinions of the meeting participants converged on using a common test linking and equating approach based on the administration of relevant common measures along with each state's own test to a sample of students. There was a consensus that the STQI Project should devote further effort to identifying and describing the conditions states would have to meet to develop a common scale by using a common test linking approach. This examination was to focus on technical considerations (timing, dimensionality characteristics of the test, sample size needed) and resource and time considerations.

In addition to the recommendation on further study of the common linking approach, the participants recommended that CSE proceed with the following tasks:

1. Complete the interviewing about state testing activities and develop a chart that characterizes these activities.

2. Continue to obtain representative reports generated by state testing programs and conduct an analysis of their content with respect to the methodology used to develop, analyze, and report data at the state level.

3. Conduct an examination of the content of state tests including analysis of both content specifications and actual items where feasible.

4. Explore further the feasibility of developing summary Consumer Report-type indicators of trends with respect to diversity of content measures, complexity of skills measured, longitudinal changes, and subgroup differences.

5. Attempt to provide resource and time estimates necessary to both pilot and fully implement the approaches judged to be fruitful to arrive at state-level education indicators.

Recommendations from the Second Policy and Technical Meeting

To implement the recommendations from first Panel meeting, several activities were carried out by CSE staff and members of the Panel. First, to obtain a clearer statement of the technical options for employing the equating and linking strategies, R. Darrell Bock, a member of the Panel, was asked to provide a memorandum describing the psychometric alternatives and the conditions necessary to implement them. This memorandum was then circulated to other Panel members for their reaction prior to the scheduled April Panel meeting. Written feedback from other Panelists was distributed along with other materials prepared for the meeting.

Second, CSE staff conducted a detailed examination of existing tests used by states. This content analysis was intended to provide a basis for judging whether there was sufficient overlap in content coverage and grade levels assessed among the states to actually implement a linking effort. It was also hoped that this activity would suggest ways to develop indicators that portray the diversity of content covered in existing state tests.

The third major CSE activity was an examination of the state reports to determine whether there was sufficient information to develop within-state trend and subgroup comparisons to serve as indicators across states. This investigation also sought to establish the degree of overlap in the scales states used to report performance and whether states collected and/or reported auxiliary information about the characteristics of their students and schools that could be used to contextualize student performance.

At the beginning of the second Panel meeting, participants received the available correspondence with respect to the Bock memorandum on technical alternatives, the draft materials from the detailed content analysis, a draft of the survey of auxiliary information collected and/or reported by states, and a draft outline for the final report. Using these materials, participants discussed the advantages and disadvantages of two alternative strategies for applying the common linking approach, namely:

1. Matched test data strategy where scores from separate administrations of the linking test (presumably NAEP) and existing state tests would be matched at the pupil level;

2. Common anchor item strategy where the linking test and the existing state test would be administered concurrently.

Two concerns needed to be addressed before a decision could be reached about how either linking strategy might be applied. First, the question of possible content of the common tests was raised. To that end, participants examined the content analysis of tests or specifications of tests from 38 responding states who were conducting testing programs as of Spring 1984. Based on these data, the panelists recommended that two or three skill areas at a single grade level be chosen for initial examinations of equating options based upon the frequency of the skill areas' inclusion in State measures and the frequency at which various grade levels were represented in State test administrations. The areas of literal comprehension in the reading achievement area and either numbers and numeration or measurement in the mathematics achievement area at grades 7 through 9 were considered most suitable for initial equating efforts.

The second concern was the nature of the common measure proposed to serve as the basis for equating the disparate state measures. It was determined that technical procedures now exist that make it possible to equate tests without requiring that all sampled students respond to the same set of common items. However, the measures needed to share certain technical characteristics with the target measures in reading and math. Principal among these characteristics was unidimensionality of the scale.

The remainder of the discussion focussed on the source of items for the common linking measure. Three alternative sources of test items received the greatest attention: NAEP, commercially available standardized achievement tests, and items from state-developed tests. The strengths and weaknesses of each of these options were explored. Among those present at the end of the meeting, a preference was expressed for drawing primarily from a pool of items developed by the states as this option best limits the federal presence and retains states' control of the linking effort. However, it was recognized that all sources could provide items that could contribute to a broad-based linking effort. It was also understood that this preferred option required substantial cooperation among states, additional burdens on state testing programs, and increased testing costs that would have to be borne by some level of government. These factors might lead the affected Federal and State agencies to prefer expanded NAEP testing despite its drawbacks if the latter could be done more cost effectively.

The Panelists felt that it would not be possible to decide whether the common linking strategy was feasible without conducting an exploratory study of the conditions that could affect the equating effort. Specifically, they recommended that the common anchor item strategy be tried on an exploratory basis for a two-year period, after which judgments about continuation, modification, or expansion could be made.

Following the Panel meeting, CSE was expected to complete their examinations of tests and reports to provide as complete a documentation as possible to inform decision-makers and persons charged with implementation of the chosen option. It was agreed at the April Panel meeting that reporting of project results was to be done at two levels. A decision memorandum describing study purpose and procedures, options considered and recommendations was to be prepared for the Director of NCES.* A larger report that provides details of all project activities was to be prepared with a broader target audience of both federal and state officials interested in current practices in state testing and their potential for contributing to comparative indicators of education quality.

Overview of the Report

This report is intended to provide the detailed documentation of the activities carried out under the auspices of the STQI Project. Given the diverse interests and expertise of its target audiences (primarily policy makers, their staffs, and state testing practitioners), we have tried to separate reporting of the main themes in the investigation from more fine-grained treatment of the details of state testing practices. Much of the latter has been relegated to appendices.

---

* A copy of the decision memorandum appears in Appendix 4. This memorandum was submitted July 30th. Subsequent to its submission, there were slight modifications in certain project recommendations in response to additional input from project panelists and state and federal officials concerned about education indicators. However, the main thrust of the final project recommendation remained consistent with the earlier memorandum.

The remainder of this report is divided into four separate chapters on specific project activities plus a summary chapter and appendices. The description of existing state testing programs is provided in Chapter 2. This chapter describes CSE's procedures for obtaining the information about programs, other sources of information about these programs, and provides cross-state summaries of current practices. In Chapter 3, alternative approaches for using a common test linking strategy for expressing state results on a common scale are considered in detail. Included in this examination are descriptions and evaluations of basic psychometric alternatives, delineation of possible sources of test items to contribute to the linking tests and implementation issues associated with the preferred options for linking. The results of the detailed content analysis of existing state tests are reported in Chapter 4. In addition to the basic facts regarding present test contents, we attempted to highlight exemplary practices and to document the choice of content areas and grade levels for the exploratory study recommended by the Panelists. The project effort in documenting reporting practices and the collection and use of auxiliary information about student and school characteristics is provided in Chapter 5. Current practices and possibilities for reporting between-state comparisons of within-state longitudinal and subgroup performance contrasts are emphasized. In addition, recommendations are made for improving state practices in the collection and reporting of auxiliary information.

While the above overview accurately characterizes the substance of our report on prevailing practices, it does little to place its contents in perspective with respect to either the forces that led to its initiation or the multitude of in-progress changes in state testing practices. As we see it, this project was initiated to inform a policy formation process wherein historically federal and state agencies have contended over the prerogatives in documenting national educational progress. At present, however, both levels of government (the federal through its annual reporting of State Education Statistics and education indicator efforts, the States through the actions of the Council of Chief State School Officers (CCSSO) endorsing cross-state comparisons and establishing a Center on Assessment and Evaluation to coordinate information on state practices and to support efforts to align state programs more closely) have initiated actions that could lead to the gathering and reporting of comparative state-level data on educational achievement.

But the basis for these comparisons, the organizational and administrative mechanisms for compiling them, and the sources of support for the necessary expansions in data collection and reporting remain to be determined. It may well be that alternatives preferred on purely technical and organizational grounds are too costly or too politically onerous for either federal or state agencies, or that cost-effective alternatives too dramatically change the balance of roles and responsibilities. In either of these circumstances, the current will to cooperate between the federal government and the States in the development of national achievement indicators could well dissolve. If this were to come to pass, it is highly unlikely that the kinds of alternatives that we were charged to investigate could ever be implemented. Whether the country would be left then with present practice (i.e., SAT/ACT comparisons) or two competing systems is unclear; neither of these alternatives would seem to be desirable.

The other major caveat that must be considered in reading this report is that current state-level reform efforts are bringing about significant changes in current state testing practices. If current plans on various state drawing boards are implemented and maintained, more students will be tested at more grade levels in a broader array of subject matters for a greater number of states. These changes could eventuate in an expanded base of commonality of testing practices and thus enhanced possibilities of using state testing data for comparative purposes.

In the short term, however, it means that attempts to document existing state practices are inherently imprecise. At various points in our investigations, we have been forced to choose between describing what existed at the time of our data collection, what was currently being implemented, and what a state anticipated would happen in the near future. The state of Mississippi is illustrative here. According to practices prior to 1984 (as reported in the Southern Regional Education Board's report on test results from the South), Mississippi operated both an assessment program which used a commercially available standardized achievement test and a minimum competency testing program. The Education Commission of the States' December 1984 report on current state assessment practices cites only the former program. Our own sources of information portrayed a mixed picture of a system in transition where a state-developed test was planned for implementation within the next three years. As a result, we classified Mississippi differently depending on the specific issue we were attempting to address. These kinds of apparent inconsistencies appear throughout the chapters of the report although as best we can determined, they have no impact on either our interpretations of the data or our study recommendations.

What the active change efforts at both federal and state levels did mean for our project was that we found it necessary to adopt certain basic guiding principles about how intrusive the options recommended could be with respect to existing practices considering what was likely to occur in the near future. That is, since both federal and state agencies are committed to cross-state comparisons and state testing programs are changing, we thought it reasonable to consider alternatives that would require greater uniformity in practice than currently exists and that depended on multi-state cooperation to develop the desired achievement comparisons. At the same time, we took our charge to concentrate on state testing data as the basis for comparative indicators to mean that the preferred options should leave as much discretion as possible to the States collectively. To achieve this desired goal while ensuring that the resulting comparisons have a firm technical base, we assumed that the following basic principles should guide our examinations of alternative approaches for deriving comparative achievement data based on existing state testing programs and practices:

1. Existing state testing procedures should be disrupted as minimally as possible. Only those data collection activities considered essential for obtaining evidence of comparability should be introduced, over and above the states' own planned expansions and extensions of their testing activities.

2. Existing state tests and testing data should be used as much as possible. Thus, to the extent that is feasible, state test data would serve the multiple purposes dictated by both its original intent and the desire for cross-state comparisons.

3. Regardless of the specificity desired in the reporting of cross-state performance, the content of the tests to be used for comparison purposes should be specified at as low a level (subskill or subdomain if possible) as possible to enhance the quality of the match of existing tests to the linking tests and to encourage attention to the details of what is being tested.

4. The content covered by the linking tests should be as broad as possible both to ensure some degree of overlap with each state's tests and to encourage broadening rather than narrowing of the curriculum across the states.

5. While the present project charge by necessity focuses discussion on state-by-state achievement indicators, the proposed approaches should be compatible with the wider issue of the development of systems for monitoring practices and progress both within and across the states. Augmentations of present state practices that encourage improvements in documenting the characteristics of its students and schools within the framework of planned changes in state educational activities at minimal added expense are desirable. To the extent possible, these augmentations should be designed to serve the dual purpose of a national monitoring system as well.

In essense, we are examining the feasibility of developing a set of state-by-state achievement indicators that grows out of existing state testing activities. The resulting set of indicators should draw heavily from the content specifications and item pools collectively administered by States but by necessity may include content unevenly distributed among current state tests. Ideally, the proposed achievement indicators should build upon and extend the capacity of individual States to monitor comparatively the progress of their students within a broad framework of curricular objectives arrived at through collective and collaborative decision-making by representatives of the States. The purpose of this project, then, is to ascertain the conditions that support or impede progress toward this ideal and where possible, to suggest feasible modifications and extensions of current testing activities to better approximate the intended goal of a national set of state-by-state achievement indicators.

Chapter 2
Description of Existing State Testing Programs

A description of existing state testing programs is presented in this chapter. CSE's procedures for obtaining information about programs are described; other sources of information about state testing programs are identified; and current practices are summarized. While this description may be of direct interest to policy makers and practitioners, its primary purpose with respect to this report is to establish the context of existing practices within which alternatives for linking test results across states must be considered. For this reason the discussion of state testing practices will be brief and will focus on information that can hopefully clarify and refine the consequences of the test linking alternatives.

Procedures

Part of the basic charge to CSE in conducting the STQI Project was to document existing state testing program activities with specific emphasis on the possibilitity of using data already routinely collected to form "comparable" state-level achievement indicators. At the start of the project, federal personnel involved in education indicators work had only limited information about current state testing activities and viewed the project as an opportunity to rectify this situation.

To complete the compilation of information about state testing programs in the limited time allotted for the effort (Originally, the STQI Project was to be carried out within a five-month period from September 1984 through January 1985. However, the project did not actually begin until October 1984 and was subsequently extended in response to changes in objectives arising out of the Panel meetings), it was decided to conduct a telephone interview survey with representatives from the testing programs in each state currently conducting such a program. A preliminary list of contact persons in each state was obtained with the assistance of the CCSSO and the state testing members from the project Panel. Attempts were made to contact a testing representative in each state; however, this was not possible in some states which do not currently operate testing programs nor have anyone designated with responsibilities in this area.

State participation. Most of the telephone interviews were conducted during the month of November 1984. By the end of the project, representatives from every state operating a statewide, state-administered testing program sometime during the 1983-85 period were contacted. In total testing representatives from 42 states were interviewed and/or supplied CSE with reports and documents pertaining to their state testing activities.

Four participating states (Mississippi, which disbanded one state testing program after 1983 and is currently implementing a new program; Indiana and Massachusetts, which are currently implementing state-administered programs for the first time; and New Hampshire, which had a program in the late 70's and is beginning a new one this year) were not administering statewide tests as of December 1984. Eight other states (Colorado, Iowa, Nebraska, North Dakota, Ohio, Oklahoma, South Dakota, and Vermont) do not currently administer statewide tests and did not provide CSE with information about their testing activities. Some of these states are either planning to conduct statewide assessments or already operate programs emphasizing voluntary participation or local choice of tests to administer as part of the program. Since our interest is in programs which uniformly administered a statewide test, further information about the programs in these states was not pursued following the initial round of telephone calls.

Focus of Interviews. Information about general characteristics of a state's testing program, the types and contents of reports prepared and distributed, and the availability of the data for further analyses beyond those the state included in its reports were collected during the telephone interview. A copy of the telephone interview guide is contained in Appendix 2. In addition copies of existing reports and content specifications generated by state testing programs were requested. The reports submitted by the states were used to clarify aspects of the information collected during the interview and to serve as a primary source for the examination of reporting practices (Chapter 5 of this report).

In designing the instrument for gathering state testing program descriptions, a primary distinction was made between "assessment" and "competency" testing programs. The actual label attached to a given state's testing program might vary, making its classification ambiguous. Assessment test results are most often used for general program monitoring and accountability within the state, primarily at the school and district levels. Typically, these tests cover a broad base of content and include items with a wide range of difficulty. Many states use commercially available standardized tests for their assessment purposes. Others develop their own tests (modeled after the original NAEP assessments in certain states).

Competency testing programs, on the other hand, typically are intended to measure whether students have acquired a set of skills ("competencies") viewed to be important for some educational or social purpose. Competency test results are most often used for decisions about grade promotion, high school graduation, early exit, and eligibility for remediation programs. The skills tested are generally drawn from a narrower content band than with assessment tests. "Basic skills" or "functional literacy" are emphasized with the expectation that most students at the grade level should have mastered the competencies being tested; hence 70 to 80 percent correct answers are usually established as the passing or mastery level on these tests.

When the state testing agency administers the competency program itself, the competency tests are usually specially developed rather than off-the-shelf achievement tests from commercial publishers. Many states operating competency testing programs, however, leave the choice of content and the selection of mastery levels to the discretion of local school districts. In these cases, there is a statewide competency testing requirement but no statewide, state-administered testing program. Results from states operating local option programs cannot be compared (through linking) with results from other states unless the tests administered in different locales within the state have first been equated. Because of these added complications, later discussions regarding the number of states whose programs could be linked exclude local option states even though our (and Pipho's (1984), for that matter) tabulations of existing programs includes them.

Some states operate both assessment and competency testing programs including a few cases where both programs are administered at the same grade level. During our interviews information about program characteristics was recorded separately for assessment and competency programs so that we are able to identify instances of multiple programs operating at a given grade in the same state.

In the descriptions that follow, special attention will be paid to program characteristics that are likely to have the greatest impact on whether a state's test data can be used in the linking effort. Of particular interest are (a) the content areas tested (reading, mathematics, writing, and other (typically language arts, social studies, and science)), (b) grade levels tested, (c) dates of test administration (Fall, Winter, Spring or actual month), (d) sampling strategy (census (every person at a grade level without a special exemption) or sample (a random or stratified random sample of students or schools), (e) sources of test items (internally developed or commercially published), and (f) indications of plans for major program changes.

Before proceeding with the discussion of results of our phone interviews, it is important to note the existence of other recent surveys of state testing activities. A list of other sources of information about these programs which we identified during the course of our investigation is contained in Appendix 5. The December 1984 reports on the current status of state assessment and minimum competency testing programs prepared by staff at the Education Commission of the States (ECS; Anderson, 1984; Pipho, 1984) and the results from the Roeber surveys of testing directors are most relevant to the current effort. In certain instances, the results of the phone interviews were combined with information from these other surveys to obtain a presumably more accurate picture of current state testing activities. However, in a few cases, there are differences in the information reported by the various surveys, most likely due to differences in when and how specific questions about program characteristics were asked. For the most part, discrepancies are minor and it should not matter which description is considered definitive.

Summary of State Testing Activities

The basic results from our examination of state testing activities are presented in a series of tables and a figure. The detailed summary of state-by-state program characteristics is reported in the table appearing as Appendix 6. Specific features of a state's assessment and competency programs are reported separately in this table. The prevalence of both types of testing activities is portrayed pictorially in Figure 2.1. In this figure local option competency programs are included only when the state also has an assessment program. State-by-state information about the dates for test administration for assessment and competency tests is provided in Table 2.1. Finally, if the distinction between assessment and competency testing is ignored, the pattern of content areas and grade levels tested across the states is as depicted in Table 2.2.

When aggregated across all states, the main characteristics of state testing activities can be summarized as follows:

1. Number of Statewide Programs -- As of December 1984, 39 states (including Mississippi) were operating at least one statewide testing program.

2. Assessment Programs -- 35 states were conducting statewide assessment programs. This number includes Mississippi (recently discontinued) and three states (Florida, Michigan, and Texas) whose programs serve both assessment and competency purposes according to state testing officials. Other states not currently conducting statewide assessments (Idaho, Massachusetts, and South Dakota, according to the ECS survey) plan to start such programs in the near future.

3. Competency Programs -- 36 states currently operate minimum competency testing (MCT) programs; 9 of these programs are local option according to our survey. (Note: The December 1984 ECS survey conducted by Pipho identified 38 states with MCT programs, excluding Colorado. However, his list does not match ours exactly. We have excluded from our list some states where the testing director did not classify the program as MCT even if Pipho did. Also, there are some states (Massachusetts, Nebraska, New Hampshire, Ohio, and Vermont) which operate local option competency programs according to Pipho but did not complete the CSE interview due to the absence of a statewide, state-administered program.

4. Multiple Programs -- 22 states operate both assessment and competency testing programs while 3 additional states use the

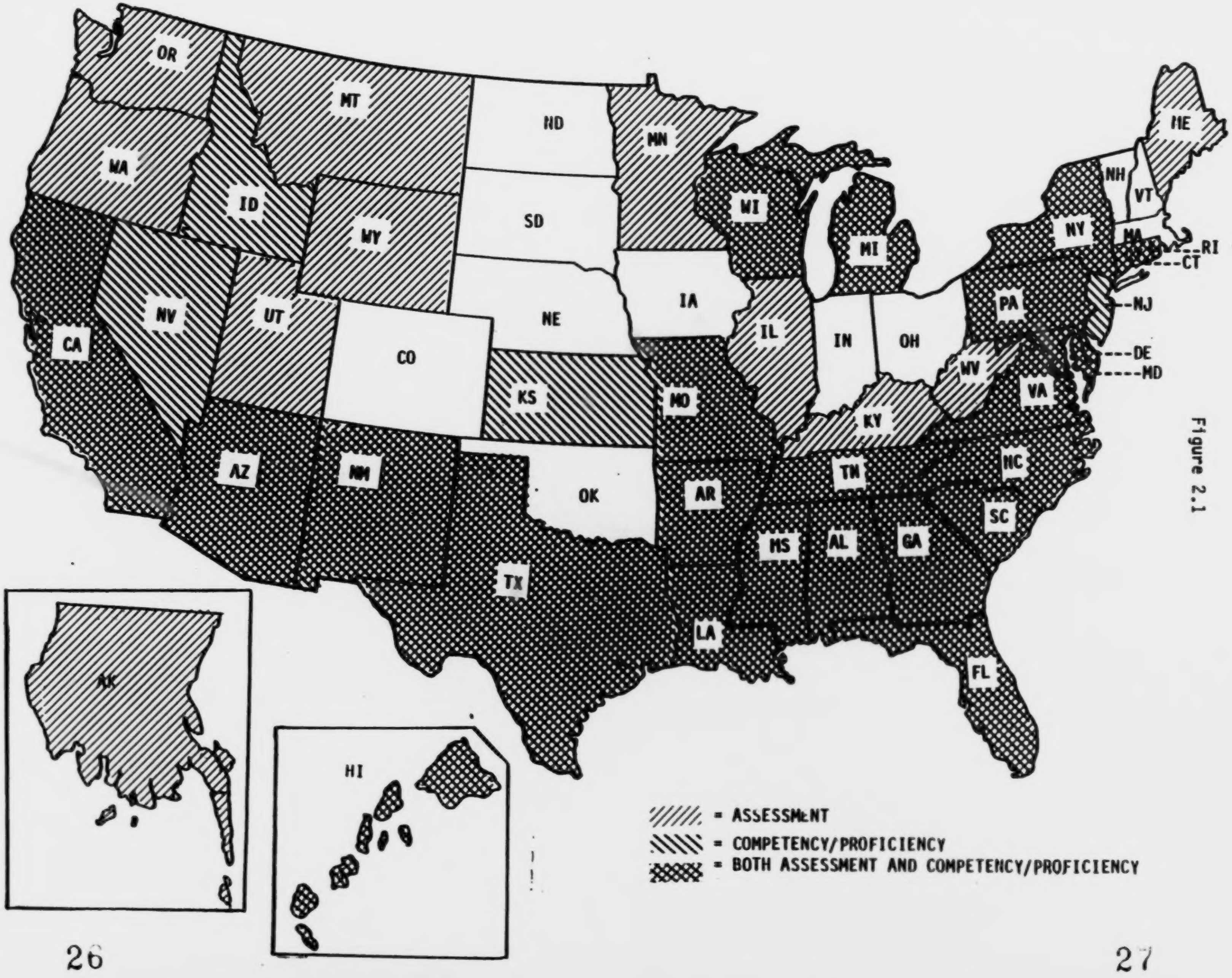

OR
MT
ND
MN
ME
WA
ID
WY
SD
WI
MI
NH
VT
NY
MA
RI
CT
NV
UT
IA
PA
NJ
CA
CO
NE
IL
IN
OH
DE
MD
WV
VA
KS
MO
KY
AZ
NM
OK
AR
TN
NC
SC
MS
AL
GA
TX
LA
FL
AK
HI
▨ = ASSESSMENT
▧ = COMPETENCY/PROFICIENCY
▦ = BOTH ASSESSMENT AND COMPETENCY/PROFICIENCY

Figure 2.1

TABLE 2.1

Administration Dates for State Testing Programs

STATE	STATE ASSESSMENT TEST DATES	COMPETENCY PROGRAM TEST DATES
Alabama	April	October (Grades 11 & 12)
Alaska	Every 2 years in March	------
Arkansas	April	April
Arizona	April	------
California	April - May	?
Connecticut	?	October
Delaware	March	?
Florida	------	March (once every 2 years)
Georgia	Spring	------
Hawaii	Fall (September-October)	Spring (May)
Idaho	------	Grade 11 - April Grade 8 - February
Illinois	Spring	------
Indiana	------	February (Starting 1985)
Kansas	------	April
Kentucky	April	-------
Louisiana	March	March
Maine	Grade 8 - Fall(late Nov.) Grade 4 - February Grade 11 - April	------
Maryland	Fall	?
Michigan	September - October	?
Michigan	Fall	Fall
Minnesota	4 - Winter, 8 - Fall 11 - Spring	------
Missouri	Fall	Fall
Montana	April	------

STATE	STATE ASSESSMENT TEST DATES	COMPETENCY PROGRAM TEST DATES
Nevada	------	Fall; Spring for Fall Failures
New Jersey	------	Spring (March)
New Mexico	March	Spring
New York	Spring	Spring
North Carolina	First Week in April	Spring - Field Test 2 x (Oct. & May) & June for Seniors Only
Oregon	Every Four Years; is going to change to every year, March	------
Pennsylvania	March - April	March - April
Rhode Island	Spring (April)	Fall (November)
South Carolina	March - April	March - April
Tennessee	Spring	?
Texas	February	February
Utah	Every Three Years in Spring (mid-April)	------
Virginia	Spring	February
Washington	Grade 4 - October Grade 8 - February Grade 11 - Late April	------
West Virginia	3-6 Spring, 9-11 Fall	------
Wisconsin	Spring	Spring
Wyoming	Spring	------

TABLE 2.2

OVERVIEW OF CONTENT TESTED BY GRADE LEVEL

CRT's & NRT's

CRT Major Content X Grade Level

Key
R = reading
M = math
W = writing
- = norm referenced test

LIST OF STATES FOR STQI PROJECT

Comments:		Grade 1-3	Grade 4-6	Grade 7-9	Grade 10-12
	ALABAMA	(CAT) RWM	(CAT) RWM	(CAT) RWM	(CAT) RWM
	ALASKA		RM	RM	
	ARIZONA	(CAT)	(CAT)	(CAT)	(CAT)
	ARKANSAS	RM	(SRA) RM	(SRA) RM	(SRA)
	CALIFORNIA	RWM	RWM	RWM	RWM
No program	COLORADO				
	CONNECTICUT	RWM	RWM	RWM	RWM
	DELAWARE	CTBS (1-3)	CTBS(4-6)	CTBS (7,8)	CTBS(11)
	FLORIDA	RWM	RWM	RWM	RWM
	GEORGIA	RM	RM	RM	RM
	HAWAII	(SAT)RWM	SAT	SAT,DAT	RWM(STAS)
	IDAHO			RWM	
	ILLINOIS	RWM		RWM	RWM
New 85	INDIANA	RWM	RWM	RWM	
No program	IOWA				
	KANSAS	RM	RM(4&6)	RM	RM
	KENTUCKY	CTBS-U	CTBS-U	CTBS-U	CTBS-U
	LOUISIANA	RWM(2,3)	RWM	RWM	RWM
	MAINE		RW	RW	RW
	MARYLAND	(CAT)	(CAT)	(CAT)RWM	
Districts choose - no statewide test	MASSACHUSETTS	RWM	RWM	RWM	
	MICHIGAN		RM	RM	RM

		Grade 1-3	Grade 4-6	Grade 7-9	Grade 10-12
Content differs by grade	MINNESOTA	M	RM	RM	RM
	MISSOURI		RWM		RWM
	MONTANA		RWM		RWM
	MISSISSIPPI	RWM	RWM	RWM	RWM
No program	NEBRASKA				
	NEVADA	SAT	SAT	RWM	RWM
	NEW HAMPSHIRE		RM	RM	RM
Local choice 3,6	NEW JERSEY			RWM	
Grade 11 = local option	NEW MEXICO	CTBS-U	CTBS-U	CTBS-U	CTBS-U
	NEW YORK	RM	RM	RWM	RWM
	NORTH CAROLINA	(CAT 1-3)	(CAT)	(CAT)	RM
No program	NORTH DAKOTA				
No program	OHIO				
No program	OKLAHOMA				
	OREGON		RWM	RWM	RWM
W = district choice	PENNSYLVANIA	RM	RWM	RWM	W
	RHODE ISLAND		ITBS(4,6)	ITBS	
No information on CRT	SOUTH CAROLINA	RM(1-3)	CTBS-U RWM	CTBS-U RWM	CTBS-U RWM
No program	SOUTH DAKOTA				
	TENNESSEE				RWM
	TEXAS	RWM	RWM	RWM	
	UTAH		CTBS-S		CTBS-
No program	VERMONT				
No information on CRT	VIRGINIA		SRA	SRA	SRA,RM
	WASHINGTON		CAT		
	WEST VIRGINIA	CTBS-U	CTBS-U	CTBS-U	CTBS-U

	Grade 1-3	Grade 4-6	Grade 7-9	Grade 10-12
WISCONSIN		CTBS-U,R	R CTBS-U	CTBS-U,R
WYOMING		NAEP	NAEP	NAEP

	Grade 1-3			Grade 4-6			Grade 7-9			Grade 10-12		
Total number of states testing R,W,M	R	W	M	R	W	M	R	W	M	R	W	M
CRT [May also do NRT]	17	11	17	23	14	22	25	19	25	24	16	22
NRT only (assumes all NRT include R,W,M)	7	7	7	12	14	13	8	10	9	6	8	7

same test for both purposes. 18 of these states administer two separate statewide testing programs.

5. Content Areas -- Virtually every state operating a program tests in the content areas of reading and mathematics. Less than half the states conduct writing assessments while over half also test in either language arts, science, social studies or some other area. In Chapter 4, we examine the content of state tests in greater detail.

6. Type of Test -- 20 states report the use of one of the major commercially published standardized achievement tests in their statewide assessment or competency testing programs. In 32 states, at least one statewide test is either internally developed (perhaps by an outside vendor according to state specifications) or involves a concurrent assessment of NAEP tests.

7. Grade Levels Tested -- Statewide testing programs are most frequently conducted in grades 8 (32 programs (T), 22 assessment (A) and 10 competency (C) with 5 states conducting both at this grade level (B)), 11 (29 (T), 16 (A), 13 (C), 3 (B)), 3 (27 (T), 14 (A), 13 (C), 2 (B)), 4 (25 (T), 18 (A), 7 (C), 3 (B)), 10 (24 (T), 12 (A), 12 (C), 4 (B)), and 6 (21 (T), 13 (A), 8 (C), 1 (B)). The fewest programs are conducted at grades 1 (8 total), 2 (11), 7 (12) and 12 (13). See Chapter 4 for further examination of grade levels tested.

8. Dates of Test Administration -- The majority of states conducting statewide testing programs administer at least one test during the Spring (typically March or April). Several states currently conducting concurrent assessments with NAEP during the Fall will shift to Spring testing when NAEP does. See Chapter 4 for further discussion of dates of test administration.

9. Type of Sampling -- At least 24 of the 35 statewide assessment programs conduct census testing in most content areas. According to our records, all statewide competency programs test every eligible student at the target grade levels.

10. Planned Changes -- Almost every state currently operating a testing program is planning a major change during the next few years (at least 36 states including those starting new programs, by our rough count). The most frequently mentioned changes are the addition of new grade levels, expansion to new content areas (direct writing assessments, science, social studies), tests of higher-order skills, change of commercial test used, redesign of program, revision of competency tests, concurrent assessment with NAEP, shift to census testing, and change in use of competency tests (e.g., adding a graduation requirement or a mastery component).

The above points highlight the substantial amount of testing activity currently being conducted by states. While there is substantial variability across states in specific program characteristics, there is some degree of convergence on content areas, grade levels, and dates of test administration. At this somewhat superficial level, then, it appears that it would be feasible to pursue further the possibility of comparing test results (through linking and equating) from a significant number of states in certain content areas at certain grade levels. Of course, the potentially serious effects of testing conditions (e.g., type of test, grade level and dates of administration differences) on the accuracy of the linking would have to be determined and taken into consideration in any comparisons.

The other major caveat that must be considered is that state testing practices are obviously undergoing significant changes in response to state-level reform efforts. It appears likely that in the near future, more students will be tested at more grade levels in a broader array of subject matters for a greater number of states. If these changes actually occur as planned, there would be an expanded base of commonality of testing practices, thereby improving possibilities of using state testing data for comparative purposes. Whether these changes will occur, and programs stabilize at this higher level of compatibility, remains to be seen.

While we will withhold making most of our recommendations until later chapters, there is at one that derives directly from the issues addressed here. Federal and State policy makers interested in the impact of state reforms will continue to need updated information about state testing activities. Regardless of whether state test data contributes to the set of national achievement indicators, these programs do change in response to reform efforts and in many cases, serve as the basis for state and local assessments of the impact of reforms. Under these circumstances, we believe that it is essential to support recurring collection of data about state testing activities that can contribute to the information base for federal and state (both individually and collectively) policy formation.

Chapter 3
Consideration of Common Test Linking Strategies

Statement of the Problem

At the heart of the STQI Project's charge was the question of whether there is some feasible way of linking existing state tests to a common scale for state-level comparisons. The information about existing practices cited in the previous section points to the crux of the problem: even given the general impetus toward expanded testing, there is still substantial diversity in state practices that presents potential obstacles for a routine, straightforward linking and equating effort. The major potential obstacles can be summarized as follows:

1. There are no statewide testing programs in some states. Eleven states (Colorado, Indiana, Iowa, Massachussets, Nebraska, New Hampshire, North Dakota, Ohio, Oklahoma, South Dakota, Vermont) do not operate either a state-administered assessment or minimum competency testing program at this time. Although several of these states are in the process of establishing statewide testing programs (Colorado, Indiana, Oklahoma, South Dakota and Vermont are in various stages of development according to our sources), there is still no test to equate in some states and probably will not be one over the next several years.

2. There is substantial variation among states in the focus of the content tested. Some states opt for broad-based assessments including direct writing assessments and the measurement of critical thinking while others concentrate on basic skills that all students at a given grade level are expected to have ("minimum competencies"); some states do both.

3. The source of the tests used for state testing varies. Some states develop their own customized tests, others choose to administer a publisher-provided standardized achievement test, and still others customize a publisher-provided standardized test. Regardless of source, some states change either the test (e.g., from one publisher-provided test to another) or modify its content (generate new items, expand content coverage) regularly.

4. States test at different grade levels. While testing is conducted in certain grades in many states (grades 8, 11, and 4, the grades covered by NAEP testing, are most popular), there is only a few grades where a majority of the states currently administer tests.

5. States test at different times of the school year with April, March, and October the most popular months. In some states selected grade levels are tested in the fall while testing is conducted during the winter or spring at other grade levels.

6. Some states exhaustively test all students at chosen grade levels while others collect data from only a sample of students at any one grade.

Obviously, if the development of an achievement indicator for comparing states requires that all states test a comparable sample of students on equivalent content at the same grade levels at the same time of year, it would be impossible to meet the conditions necessary to establish such indicators in the short term. This is the case despite the professed federal and state interests in developing a better set of achievement indicators and a willingness to explore state-based options as data sources.

The short-term picture (and presumably the long-term situation as well) is less dismal if it is not essential that all states be included in the comparisons and the other conditions for comparability are relaxed. The basis for relaxing the conditions should be that the comparison of the performance of states should only be made if there is sufficient empirical evidence to allow analytical adjustments for the effects of differences in administration conditions. Thus, even if State A normally tests a sample of its students using a minimum competency test at grade 7 in the fall and the chosen target grade and date for comparison is grade 8 in the spring, State A's performance can be compared with performance in other states if the effects of the differences in that state's testing conditions can be ascertained and a reliable and valid means for making the necessary adjustments is available. The effort necessary to obtain this evidence could be substantial, but the problems are more with logistics (obtaining the necessary cooperation and conducting the necessary special studies) and economics (obtaining the required funding for the special studies) than with technology. The methodology for generating the actual adjustments and incorporating them in the comparisons is well-established with the most difficult part being to determine all the conditions that need to be empirically investigated.

In the remainder of this section, we set aside for the moment questions about whether all states conduct testing programs and substantive concerns about the actual content of tests in order to focus attention on the alternative analytical approaches for expressing the test results from different states on a common scale. This examination will concentrate on logistical details of the psychometric alternatives considered rather than on the psychometric details themselves. Moreover, the focus will be on a few alternatives that the STQI Policy and Technical Panel viewed to be of greatest potential interest.

Procedures for Examining Alternative Approaches

At the November 1984 meeting of the STQI Project Policy and Technical Panel, a number of alternatives were considered for arriving at achievement indicators from existing state testing activities. The summary of that meeting provides details of the discussion and is partially reproduced in Appendix 3. The project charge following the meeting was to concentrate on elaborating the procedures for using equating and linking methodologies for arriving at a common scale for cross-state comparisons. Specifically, what additional new data collection would be necessary to apply these approaches in a substantial number of states and what are reasonable time and cost estimates for their expanded, full implementation?

The Panel's recommendation on further examination of the equating and linking strategies was implemented by asking (a) Darrell Bock to provide a memorandum describing the psychometric alternatives and the conditions necessary to implement them and (b) other members of the Panel to react to Bock's memorandum prior to their April 1985 meeting (A number of Panel members provided written feedback following this meeting). In addition, CSE staff were to conduct a detailed examination of existing tests used by the states to provide a basis for judging whether there was sufficient overlap in content coverage and grade levels assessed among the states to actually implement any linking strategy of existing state tests.

The results of these two activities (the Bock memorandum plus Panelists comments (See Appendix 7) and the detailed content analysis of existing state tests) served as a starting point for an extended discussion of the strengths and weaknesses of various alternative approaches at the April 1985 meeting of the Panel. At the conclusion of the April meeting, the consensus among the panelists present was that

- o A pilot study of selected variations of one approach (the common test linking strategy) should be conducted in a limited set of skill areas for a specific grade range in order to determine both the quality of the equating under preferred conditions and the effects of various deviations from these conditions.

Basic Psychometric Alternatives

Stripped of details about the content to be scaled across the states, and the source of items to serve as a link, there are two basic psychometric alternatives for placing state test results on a common scale that would involve existing state tests (in contrast to the conduct of expanded NAEP testing):

1. Matching scores from the test (items) chosen to serve as a link with existing state test scores (matched test data)

2. Concurrent administration of the linking test and the existing state test (common anchor items)

Both alternatives would require that a "common linking test" be administered within participating states to a sample of students and schools of sufficient size to carry out the desired equating to a common scale.

Matched test data. The matched test data strategy would require that within a participating state, a sample of pupils be identified whose item responses to both the common linking test and the state test to be scaled could be matched. These two tests need not be administered at the same time within the state, but the ability to match at the item level for pupils is essential.

If NAEP were to serve as the common linking test, this matching would entail using the sampled schools' rosters of students taking NAEP to link student data from the NAEP public use tapes with the data for corresponding students from the state testing program. Once a sample satisfying the matching conditions has been obtained, item response theoretic (IRT) scaling methods based on marginal maximimum likelihood procedures would be employed to estimate item parameters for the state test using the parameter estimates from the common linking tests, and then the estimated item parameters for items from the state tests would be used to compute scores for pupils in the state samples. (The Bock memorandum describes the essential technical features for the scaling but the reader is referred to two other Bock references [Bock & Aitkin, 1981; Bock & Mislevy, 1982] for more complete specification of the psychometric basis for the scaling.) The resulting pupil scores (and hence their weighted or unweighted averages) are expressed on a scale that will be comparable to the scales for other states who use the common linking test.

There are several critical logistical matters that are essential to attempts to employ the matched test data strategy. Possible difficulties in obtaining enough pupils in a participating state who could potentially be matched and in securing the local school site cooperation and support for carrying out the physical matching are the most salient questions. According to ETS sources, only seven states (California,Florida, Illinois, Massachusetts, Michigan, New York and Texas) have as many as 1000 students taking NAEP as part of its standard sample. In addition, there are other states (Connecticut, Minnesota, Wyoming) who participate in a concurrent assessment using NAEP items and whose results could presumably be directly scaled to the common scale chosen for state comparisons.

Even in states with sufficient samples but whose state tests are administered at different grade levels and different times of the year from NAEP, states would have to arrange special administrations of their tests in the schools and at the grade levels of NAEP testing. In those states where NAEP samples are too small or the existing NAEP samples don't match up well with the schools and students sampled in the state's testing program (in sample as opposed to census testing states), data collection would have to be augmented (denser NAEP testing when the problem is insufficient NAEP sampling; expanded state or NAEP testing where the problem is insufficient sample match). The costs for this additional testing would have to be borne by some agency.

Under current procedures for documentation of NAEP samples, the roster of pupil names matched with NAEP case numbers never leave the local school sites. Unless the schools (or NAEP) are willing to provide these rosters to the state testing program, the actual match of student data from the two tests would have to be carried out by the local school's personnel. This requirement could introduce significant noise to the data due to recording errors, a likely occurence under these conditions where the local personnel have little stake in the accuracy of the information they are requested to provide (e.g., Keesling, 1985; Neigher & Fishman, 1985). These kinds of recording errors are not restricted to the NAEP situation; they can be expected to occur as long as the information to be recorded is of limited value to the persons expected to compile it. On the other hand, there would be no incentives to falsify information either so that intentional misrepresentations should not be a problem.

There are clearly specific obstacles to using NAEP as the common linking test in the matched test data strategy. There are other alternative testing activities that are carried out in a sufficient number of states to warrant consideration as the common linking test (e.g., the SAT,ACT, ASVAB, commercially available standardized achievement tests). But each choice introduces its own set of logistical hurdles without even considering whether the content of the tests represented by the other choices is appropriate for the desired linking.

Our analysis of the potential for the matched test data strategy for scaling purposes is that despite its theoretical promise, there are currently either insufficient data for matching in a significant number of states or the existing practices with respect to the proposed common linking test (whether one chooses NAEP, ACT, SAT, ASVAB, CTBS, etc.) would have to be modified to reduce the logistical and economic burdens they would entail. Moreover, there is a feasible alternative that takes advantage of the same psychometric methodology and requires substantially less effort and expense at the lower organizational levels of the educational system.

Common Anchor Items. The common anchor items strategy requires that a set of anchor items be administered concurrently with all state tests that are to be linked. The same item response theory methods for expressing scores on a common scale that were described as part of the matched test data strategy are applicable here as well. The main distinction between the two strategies is that here the linking test is incorporated into the state's regular testing (either through embedding items or adding the anchor items to the beginning or end), thereby placing the data collection burden upon the states rather than on the local school sites. In those states which currently manage their own data collection activities, the logistics would be simplified and the reporting and recording errors would presumably be no greater for the anchor items than they are for the state's own test.

States that do not currently conduct an assessment could choose to administer the common anchor items at the target grade levels and dates without the necessity of further equating and scaling. In states that routinely test at grade levels and times different from the target grades and dates, special administrations of the common anchor items (and preferably the state's own test as well) would have to be arranged along with the collection of the anchor items at the time of the normal state test data collection. These special administrations would be needed to provide the data to determine whether there are grade level and date-of-testing effects that warrant adjustment.

The methodology to be used for equating the state tests with the common linking test does not require that all students taking the state test also take the linking test or that all students taking the linking test take the same set of items. The sample of students taking a test item from the common anchor set must be large enough to estimate the scaling constants for the state test items directly from the item responses without having to calculate individual student test scores (See Bock memorandum in Appendix 7 and referenced papers.) The important size factor is schools rather than students. Bock estimates that approximately 40-50 schools would have to be sampled at each grade level to adequately represent the population in most states for scaling purposes.

The items from the common anchor set can be matrix sampled; that is, students could take different subsets of the test items from the common anchor item pool. This testing design has been used by NAEP and many states to expand the sample of items from a specific content area and thus could allow more content areas to be incorporated in the anchor set for the same length test. This item sampling strategy requires more students from a given school be tested but reduces testing time in a given content area for any participating student.

The remaining logistics and consequences of incorporating anchor items into data collection in states that develop their own tests is relatively straightforward. In states using commercially available standardized tests (The CTBS, CAT, SAT, ITBS, and SRA tests are each used in multiple states at some grade levels.), there are both potential additional constraints and possible economies. If a state wished to use a publisher's tests for its standard purposes (other than for the indicator activity), the anchor items should not be seeded within the test or administered at the beginning of testing because the non-standard administration can affect the validity of the test norms. Thus the procedures for joint administration would likely be more limited in states using published standardized tests. At the same time, as long as the different states using a specific standardized test do so under the same conditions (same grade levels and time of year), it would not be necessary to estimate the scaling constants anew for every state, at least for technical reasons.

Preferred Option. Our analysis suggests that the common anchor item strategy is preferable to the matched test data strategy if a common test linking approach is to be used to express the test scores from different states on a common scale for comparison purposes. The basis for the choice is primarily logistical; the operation could be managed by the state testing agency as part of its regular testing activities without requiring potentially extensive new assistance from local schools and introducing the technical complexities of carrying out the required matching. On virtually any other aspect of the technical and logistical requirements for arriving at comparably scaled state test results, the problems are essentially the same for both matched test data and common anchor item strategies.

The common anchor item strategy places the burden for carrying out new testing activities on the state-level testing operation. The increment in effort can be large or small depending on how far the state's current testing programs diverge from the targeted testing conditions for the linking effort. This burden will also fall disproportionately on smaller states who develop their own tests and on states that change the content of their test frequently (new scalings are required for each new state item pool). If the states are to be responsible for both gathering anchor data and conducting the psychometric analyses required to express their scores on the common scale, additional technical expertise might be needed or a mechanism for obtaining technical assistance in carrying out these activities will need to be developed. Thus the common anchor item strategy could be expected to significantly impact the operation of state-level testing programs and increase their costs. While there would most likely be secondary benefits associated with the enhanced expertise from participation in the multi-state linking effort, it remains to be seen whether state testing operations will accrue direct benefits commensurate with their additional responsibilities.

Source of Common Anchor Items

To this point we have avoided addressing the thorny question of the source of the items that would serve as the common anchor for scaling the different state tests. This is not a strictly technical matter since as our content analyses (see Chapter 4) indicated, virtually all existing state-developed tests, standardized achievement tests, as well as NAEP, contain test items covering some of the skill areas that would be desirable to include in the common anchor. But each of these choices (state-developed items,standardized tests, NAEP) have different sets of strengths and weaknesses affecting their suitability for inclusion in the anchor set. There are also other sources, depending on the target content areas, for the achievement indicators; of course, new items could be written directly to fill desired content domains. Below we consider the strengths and weaknesses of the three main sources, explore the advisability of drawing upon yet other sources and provide a recommended decision strategy to select a source or sources for the common anchor.

NAEP. The test items developed for and previously administered by NAEP represent a natural pool from which to select items for the linking effort. Historically, few within the testing community have quarreled with NAEP's item writing expertise. The actual NAEP test items are of high quality and through their inclusion in previous test administrations, have associated normative data about their empirical properties. In fact, in terms of their national representativeness, the norms for previously administered individual NAEP items are probably superior to the norms of items from either commercially available standardized tests or existing state-developed test items.

Most of the limitations that NAEP would have as a linking device in the matched test data strategy (periodicity of assessment, small state-level samples, constraints on student identifiability) are no longer at issue when the question is whether NAEP items could contribute to a common anchor set. Even the supposed thinness in the content sampling of certain item domains is of less concern as long as there are other item sources that could be used to augment NAEP. The one potential technical limitation that still could diminish the value of NAEP items as a source would be the lengthy time interval between administrations in some content areas (affecting the utility of the normative information from regular NAEP administrations).

Given the availability of normative data on test items of good quality and the presumed credibility of NAEP to various stakeholders, it is sensible to include NAEP among the sources for the common anchor items. At the same time, there are reasons for incorporating items besides NAEP in the anchor set. Technically, some states have argued over the years that NAEP does not adequately reflect their own curriculum (See Roeber letter in Appendix 7). The evidence from our content analyses of existing state tests supports this contention to a certain degree, assuming that state tests cover only what is part of or should be part of the state's curriculum. There are obvious remedies to this presumed deficiency which we consider below.

Political considerations are also an important element in the argument against using NAEP as the sole source of common anchor items. Despite the extensive professional and practitioner involvement in the development of NAEP, it is, in the final analysis, a federal enterprise thus raising the attendant concerns about a national curriculum. In fact, using only items from NAEP in the anchor set would make it the national standard for comparing states in much the same way as would a direct comparison of states with an expanded NAEP with larger state samples would. The only differences between expanded NAEP and the common test linking strategy with NAEP as the sole source of items would be how items were selected (presumably some group representing states would have a major role in item selection under the common test linking strategy), the added value/complications/costs associated with the equating and common scaling, and the distribution of logistical and financial burdens for conducting the data collection. Essentially, the states, though claiming the prerogative of defining the content of tests by which they would be compared, would be virtually abdicating to a federal entity (NAEP) the actual basis for comparison. While there may be short-term technical and political advantages to such a decision, the precedent it establishes may have adverse long-term consequences for the demarcation of federal and state roles in education indicator efforts.

Commercially available standardized tests. There are several commercially available standardized test batteries (CTBS, CAT, MAT, SAT, SRA ,ITBS) that could be used as a source for the common anchor items. All of these tests have publisher-developed national norms and all sample broadly from what the publishers perceive to be national-consensus objectives (as determined primarily by textbook examinations). A significant number of states already use one of these tests as their state assessment and many districts within states who develop their own assessment tests also administer a standardized test for their own purposes (e.g., for compensatory education evaluations).

The problems with using a standardized test as the common anchor have to do with matters of test selection, test security, and the representativeness of test norms and content. Selecting a single test battery from among those commercially available would create a marketing advantage for the selected publisher and would presumably entail untoward governmental intrusion into a competitive private enterprise. The widespread use of existing batteries creates test security problems that have led to gradual deterioration of the validity of these tests as measures of learning (as opposed to test coaching) in the past; a secure form of the standardized test would be needed if were to serve as an anchor over time. The concerns about norm representativeness have to do with the problems of selective school district cooperation in publishers' norming studies (e.g.,Baglin, 1981); as a result none of the publishers have truly national norms but rather publisher-specific norms. Finally, the challenges to the contents of standardized tests have to do with their failure to incorporate important content objectives, especially at the lower and upper ends of the subject matter continuum. The traditional psychometric procedures for standardized test development select highly discriminating items that are likely to fall in the middle range of difficulty; thus content known by either most students or only a few students is typically eliminated.

These problems with commercially developed standardized tests argue against their use as the single source of common anchor items. Whether selected items from standardized tests could be included as part of the anchor is unclear. Certainly, these tests contain items covering some of the content that should be included in the anchor set and there should be substantial data about their actual empirical properties. But the tests, and hence their items, are in the private domain and publishers would have to be willing to cooperate in releasing selected items to the linking effort. Whether marketing forces would support or hinder such cooperation is unclear at the present.

State-developed items. Our content analysis of existing state-developed assessment and minimum competency tests (see next chapter) identified a wide range of both skills assessed and the quality of the test items used to measure them. Some states have been particularly innovative and exemplary in measuring selected objectives; several states (primarily assessment as opposed to minimum competency states) devote significant portions of their test content to what are normally characterized as higher-order or higher-level skills (e.g., inferential comprehension in reading with passages from different subject matters, explanations and problem solving in mathematics). In yet other states, items assessing functional literacy skills are particularly well-developed.

Taken as a whole, the set of items developed by states measure virtually every conceivable skill that one might consider to be pertinent to a comprehensive representation of the content domains of reading, mathematics, and writing. While we did not explicitly examine other content areas (e.g., science, social studies), our sense is that testing practices in these areas are also of good quality and are as broadly representative of desirable content as most other sources under consideration.

One obvious limitation of state-developed test items is the lack of nationally representative normative data in most instances. In most states, however, there is no shortage of evidence about the empirical behavior of items used repeatedly over the years of the assessment. After all, certain states annually test every student at a given grade level, yielding tens of thousands of cases for every year a test item is used. Moreover, just as with NAEP and with commerically developed tests, the items selected for inclusion in state assessments undergo multiple rounds of expert and practitioner review and empirical examination before their actual use. In addition, some states have carried out studies to equate their assessments to commercially available tests to provide national perspectives on their students' performance. So while the empirical evidence from state-developed test items differs from the evidence available on NAEP and commercially developed tests, there is no evidence of uniformly poor quality or lack of representativeness of important content and some evidence of collective broader scope.

There are political advantages in using state-developed test items as a source for the common anchor items. If the common anchor items were chosen solely from state-developed tests, the the specter of a federal presence in the specification of the basis for state comparisons could be virtually eliminated. Any option for selecting the common anchor item set that includes a substantial state role in the specification of the content to be measured and significant state representation among the items selected would provide safeguards against perceptions of federal intrusion upon state prerogatives.

There are political disadvantages as well in using state-developed test items as the common anchor core. Without any other sources of nationally normative data initially, it would take time to establish a basis for comparison (i.e., what are significant differences among states at a given point and over time) and efforts made to establish public credibility and understanding of the meaning of the comparisons. A potential additional trouble spot could be the uneven representation across states in their contribution of items to the common anchor set. States without assessments could not contribute at all while those states using commercially developed tests would have to obtain special permission before contributing. It is also clear that differences in value preferences among states would have to be overcome in arriving at consensus on which skill areas to include and which items to select. Just as with other organizations, the "not invented here" syndrome is likely to be present in certain states and will have to be dealt with.

On balance, we can see no reason flatly to exclude state-developed test items from the common anchor set and both technical and political advantages to their inclusion as a source along with other options. Technically, the basic evidence to support the inclusion of any specific state's test item in the anchor set should be the same as with any item from other sources.The logistics of data collection using state-developed items as part of the common anchor are no different from other options. Finally, the political advantages are potentially substantial while the possible political liabilities for the federal government are limited.

Other sources.It seems to us that all sources of well-developed test items with sufficient data about their empirical properties could conceivably contribute to the common anchor item set. There are test item banks operated by commercial vendors or developed by federal research laboratories or school districts that could be considered. If it were deemed important and if necessary licensing arrangements could be made at reasonable costs, items developed for the ACT and SAT could be included. There are also special purpose testing programs (e.g.,ASVAB) operated by other federal and state agencies that could serve as sources.

A particularly appealing source of potential items are those from tests used in the series of cross-national achievement surveys conducted under the auspices of the International Association for the Study of Educational Achievement (IEA). During the early part of the 1980's, studies in the content areas of mathematics (the Second International Mathematics Study), science (the Second International Science Study), and writing (The Written Composition Study) have been conducted in over twenty countries. The student performance data from these studies is nationally representative (to a greater or less degree) in most countries including several of our major economic competitors (e.g., Japan in mathematics and science, several major Western European countries). There appears to be substantial interest at both state and federal levels and from the private sector in international educational statistics and comparisons (The level of involvement of these constituencies in the April 1985 NCES-sponsored conference on international education statistics is offered in support of this inference on our part). The actual inclusion of selected items from these international studies within the common anchor would provide a beginning, although limited, opportunity for regularly collecting performance information that could be used for international as well as national comparison purposes.

Preferred option. Given a decision to proceed with the common anchor item strategy as we have recommended, our analysis suggests that *the items contributing to the common anchor set should be selected from multiple sources (NAEP, commercially available tests, state-developed test items, policy relevant and technically adequate additional sources such as the IEA tests*). There are multiple sources of items that on purely technical grounds could contribute to the common anchor item core. Both technical and political considerations lend support for selecting an anchor set that includes items from multiple sources, at least one of which is the combined pool of state-developed test items from existing testing programs. If properly implemented, the multiple sources option strikes a desirable balance among state and federal (and possibly private sector) contribution, among various normative bases for comparison once the linking has been established, and among forms of legitimation and credibility by potentially competing constituencies (the public, media, industry, and various groups representing education professionals and political interests).

While an eclectic mixture of sources is desirable, we believe that the mechanisms for establishing the skills to be included in the core, selecting items to represent the skills and specifying the rules of and acceptable for participation by individual states should be developed and administered primarily by collective representation of the states (such as through the new CCSSO Assessment and Evaluation Coordinating Center). Given the traditional state responsibility for education, significant state involvement in these phases of achievement indicator development is essential. And, as long as legitimate federal needs for achievement indicators for monitoring purposes are met, the federal presence under this proposed operation could remain benign, contributing substantively at the states'initiative and serving as a source of technical and economic assistance where appropriate.

Implementation Issues

If a decision is reached to proceed to develop a state-level comparisons using the common anchor item strategy, what additional decisions would be necessary to implement the preferred states-coordinated development of the achievement indicators? This question raises the necessary implementation issues, both with respect to the operation of the coordination of the equating effort and individual state's participation in the comparison. We are not attempting to substitute our judgments for those persons who presumably would be designated by the states to coordinate the effort and those individuals within states who would be expected to implement the activities necessary for test equating and scaling. Our purpose is strictly to point out some of the issues that the federal government, the coordinating state agency, and the states might consider if they choose to implement the proposed plan.

1. Documentation-- Procedures for documenting contents of existing state tests should be specified so that questions of what is being equated to what can be addressed.

2. Content Specification-- Specification of content represented in common anchor set should be at the lowest level possible (subskill level) even if achievement indicators, at least initially, are to be reported at higher levels (skill or content area). This level of specification minimizes the possibility of overlooking meaningful content, maximizes the possibility that selected items for the common anchor will be scalable and unidimensional, and places the greatest constraints on agreement about content assignment.

3. Criteria for Item Consideration--The minimum criteria for considering an item for inclusion in the common anchor item set should be that

- The item should measure a skill that should be represented in the common anchor item set, and
- There should be sufficient empirical evidence available about the item to ascertain its behavior for the major segments of the student population with which it will be used.

4. Item Selection Procedure-- The selection of items to represent skills in the common anchor item set should be made by teams of curriculum and testing specialists from a broad-based pool of items with as little identification information as to source as is technically feasible (to guard against political and social biases in selection). Empirical data should initially be provided without the identifying features of norm source. In later phases, additional technical information about norm quality should be considered if too many items are acceptable by other judgmental criteria.

5.Testing Conditions Specifications-- The following set of testing conditions should be specified:

- Target grades and range of testing dates should be specified along with requirements for special studies in those states who normally test outside the chosen range or do not test at present but decide to participate.
- Procedures for concurrent administration of the common anchor item set with existing state test should be specified for the various alternative types of state tests (matrix sampled, state-developed single form, commercially developed standardized test).
- Auxiliary information for checking subgroup bias and determining sample representativeness (for equating and scaling purposes) should be specified.
- Minimum sample sizes (for both schools and students) should be established.

6.Pilot Study of Testing Conditions -- A design for a pilot study of effects of deviations from target testing conditions should be developed.

Our remaining recommendations regarding the implementation of the common test linking strategy have to do with the establishment of an effective political, institutional, and economic environment for this indicator effort. First, it will be a serious matter to develop the necessary levels of political support for this activity. Key participants are, of course, the Chief State School Officers, their staffs,and other State education officials, but other prominent State officials, including the Governor, Members of Congress, and State legislators may need to be involved. Representation of members of large city school districts, the education associations and from the private sector should be participants as appropriate. Broad based support for the idea should be developed.

Second, the matter of developing an institutional structure for the conduct of this activity should be considered. The benefit of having an organization of States manage the process will avoid the specter of Federal directive, and the Council of Chief State School Officers' Assessment and Evaluation Coordinating Center proposal deserves consideration for this purpose.

Third, it is essential that technical assistance and oversight be established to assure the quality of technical and methodological operation of the linking and equating, of the content of measures, and of validity of interpretations. This oversight should be provided by a panel, perhaps modeled on the panels advising the NAEP activity.

Fourth, a long-term, secure basis of financial support for this activity should be assured. The costs will not be high but resources should be regularly available.

Summary and Recommendations

In this chapter we considered directly the alternatives for linking existing state tests to a common scale for state-level comparisons. The existing testing conditions in states that might aid or hinder the linking effort were discussed. The relative merits of two psychometric alternatives (a matched test data strategy and a common anchor item set) for linking state tests through equating to a common scale were considered in detail. Possible sources of items to serve as the common link were identified and evaluated. Implementation issues that should be addressed if a decision were made to proceed with the linking effort were delineated.

The primary recommendation was that the test linking strategy be tried on an exploratory basis (for perhaps a two-year period) after which judgments about continuation, modification, or expansion could be made. The guiding features of this exploration should be that

o The comparison of the performance of states should only be made if there is sufficient empirical evidence to allow analytical adjustments for the effects of differences in administration conditions. The exploratory study should generate this needed empirical evidence.

o The common anchor item strategy, wherein a common set of linking items is administered concurrently with the existing state test to an "equating-size" sample of schools and students, should be used as the basis for expressing test scores from different states on a common scale for comparison purposes.

o The items contributing to the common anchor set should be selected from multiple sources including NAEP, existing state-developed tests, commercially available tests, and other policy relevant and technically adequate sources such as the IEA tests.

Chapter 4
Content Analysis Of Existing State Tests

Statement of the Problem

Two of the recommendations for further work that were made at the First Panel meeting had to do with obtaining additional details about the contents of existing state tests. Specifically, CSE staff were asked to proceed with the following two tasks:

1. Conduct an examination of the content of existing state tests including analysis of both content specifications and actual items where feasible.

2. Explore further the feasibility of developing summary indicators of trends with respect to diversity of content measures and complexities of skills measured.

The impetus for these recommendations was the realization that there is little extant information about the specific content contained in state-administered tests, especially those that are internally developed. Several Panelists pointed out that not all states operating internally developed programs were conscientious about developing and publishing content specifications for the generation of test items. In addition, the match of test items to specifications and the distribution of items among objectives may be uneven in some states.

The Panelists had two specific interests for urging that more detail information be gathered about the content of the state tests. First, the psychometric technology (essentially item response theory methods using marginal maximum likelihood estimation procedures) that would be used to estimate the item parameters needed for the equating and scaling of state tests via a common linking measure require that the items to be scaled form a homogeneous, unidimensional set. This requirement typically entails that test items be scaled at the subskill (e.g., computation of percent) or skill (e.g. numbers and numeration) level (technically, Bock calls this level of classification "indivisible curricular elements") even when the indicator is to be reported at a general content area level (e.g., mathematics). Thus, details of the contents of the state tests are necessary for assigning items to homogeneous clusters suitable for linking.

Second, the question of whether there are significant differences in the content tested across states is a matter of policy interest, in and of itself. Certainly, states administer

---

Pamela Aschbacher designed and carried out the detailed content analyses reported in this chapter and prepared the description of procedures.

tests that are designed to serve different purposes (basic skills, minimum competencies, proficiency, critical thinking, higher order skills) and hence presumably cover different content. Given the widespread interest in strengthening the curriculum across the states, and the explicit or implicit relationship between what's tested and what's taught, questions about the diversity of content coverage across states become salient. This is especially likely if indicators of content coverage can be tracked over time to see their relationship to curricular changes and changes in test performance.

A caveat is in order before proceeding to describe and discuss the results of our extended content analyses. CSE attempted to examine the content of state testing programs to the extent possible within the time and resource constraints governing the project. The original strategy was to sample a few states who developed their own tests and carry out an in-depth examination of the tests' content.

As the task developed, however, it became clear that the overall goals of the project would best be served by casting the net as broadly as possible to cover as many states at as many grade levels as we could gather sufficient information to warrant a content examination. Moreover, we decided to examine commercially available standardized tests used in state testing programs as well (when we could obtain them). Because the detailed content focus was not salient at the time of the telephone interviews with state test directors, we had not specially emphasized submission of tests and content specifications in our requests for reports prepared by states. Therefore, the availability of this type of information was spotty initially although we later requested additional reports from some states.

Our efforts in this area mushroomed. By the time of the Second Panel Meeting in April, much of the detailed descriptions of state tests (reported in Appendices 15 through 18 along with the procedures for the conduct of the content analysis) had been completed. At that meeting, however, the Panelists devoted their attention to addressing the question of which option for state linking was most feasible and to specifying the parameters for a possible exploratory study of this option. While the results of the content analysis were of interest and useful for addressing the broader purposes of depicting coverage and facilitating the development of indicators of content coverage, there was actually too much detail for serving the more narrow purpose of selecting grade levels and skills to be included in the exploratory study. Rather than proceeding with further detailed work on content coverage indicators, CSE staff, instead, were urged to develop simplified depictions of the results of the content analysis to facilitate the choice of content that would be piloted.

Following the Second Panel Meeting, CSE staff worked to respond to the modified charge in the area of content examination. Much of the detailed descriptions of procedures and results of the content analysis are contained in this report. But the primary emphasis in discussing the results of the analysis will be on the simplified data presentation, resulting recommendations about target content areas for the exploratory study, and a characterization of the implications of these recommendations for state participation in the exploratory study. Further exploration of other issues is left for another study.

Procedures

The purpose of this part of the STQI Project was to examine the statewide testing programs in all the states in the content areas of reading, math, and writing in grades 1 through 12, in order to present a national picture of what is currently being done and to make policy recommendations regarding the feasibility of quality indicators in the area of content coverage.

In order to accomplish this purpose, during the brief telephone interviews conducted by CSE staff, the directors of state testing programs were requested to send CSE a copy of the appropriate tests, manuals, technical reports, and so forth. (See Appendix 8 for a list of documents provided by states.)

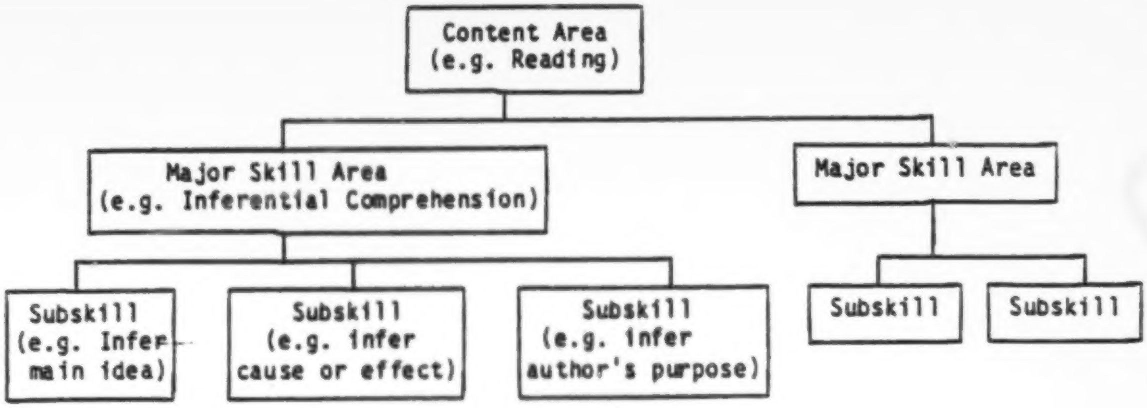
Tests included in the analysis were all currently used statewide tests given in grades 1-12 in reading, math and writing (including writing samples and writing skills such as punctuation, grammar, word usage, and organization.) The tests included those labeled assessment tests, minimum competency or proficiency exams, and inventories of basic skills. Some were commercially developed; others were criterion-referenced tests developed by state testing committees comprised of curriculum and evaluation specialists and teachers.

The analysis of tests and materials proceeded in the following manner. The objective of the analysis was to describe the breadth and depth of each state's testing program in reading, math, and writing. In order to accomplish this, "breadth" and "depth" were defined, and a matrix of major-skill-areas-by-cognitive-hierarchy was developed for each of the 3 content areas (reading, math, and writing). See Appendix 9 for these matrices.

The major skill areas and their subskills within the content areas were identified with the aid of several states' materials and three booklets:

National Assessment of Education Progress, Reading Objectives 1983-84 Assessment
National Assessment of Education Progress, Math Objectives 1981-82 Assessment
National Assessment of Education Progress, Writing Objectives 1983-84 Assessment

Content Areas, Major Skill Ares,and Subskills are related as follows:


Content Area
(e.g. Reading)
Major Skill Area
(e.g. Inferential Comprehension)
Major Skill Area
Subskill
(e.g. Infer main idea)
Subskill
(e.g. infer cause or effect)
Subskill
(e.g. infer author's purpose)
Subskill
Subskill

The major skill areas in each content area follow:

READING: (Content Area)
- Word Attack
- Vocabulary
- Literal Comprehension
- Inferential Comprehension
- Study Skills
- Attitude Toward Reading

MATHEMATICS
- Numbers and Numeration
- Variables & Relationships
- Geometry
- Measurement
- Statistics & Probabiity
- Computers, Calculators, Technology
- Attitude Toward Math

WRITING
- Conventions
- Grammar
- Word Usage
- Organization
- Attitude Toward Writing

Next, these lists of skills and their subskills (e.g. "identified word meaning in context" is a subskill of the skill area "Vocabulary) were classified according to a 4-level modification of Bloom's taxonomy of educational objectives to form the 3 content-by-hierarchy matrices. The 4 hierarchy levels included in this study were: recall, routine manipulation of literal comprehension; inference, translation, explanation, or judgement; and application, problem solving.

The materials for each state were carefully examined to classify the test items according to the content-by-hierarchy matrices. In some states, more than one test was used,so all tests of the relevant content were analyzed. The number of test items for each subskill in the matrices were recorded for each test at each grade level. For writing samples, the number and type of writing samples were recorded together with information about the type of scoring system used.

The materials received from the states varied greatly in scope and detail provided. Where actual tests were provided, they served as the primary source of data. In other cases, manuals or reports had to be relied upon to provide the information. At one end of the continuum were reports that made vague mention of a few of the skills tested but gave no comprehensive list of skills or details on how many items of each were used. At the other end were reports that included complete test specifications with detailed descriptions of objectives, skills, sample test items, and number of such items on the tests by grade level.

For each state, CSE staff attempted to extract the most specific level of data possible. Hence, for some states it was only possible to indicate that certain subskills were indeed tested without any indication of the number of such items on the test. For others, it was difficult to match their descriptions of the test content with the matrices of subskills for several reasons, often because some of the test reports lumped several different subskills together with only a total number of test items specified or because the reports gave overly brief descriptions of the skills tested (e.g. "main idea" did not specify whether the student had to identify an explicitly stated main idea or infer it from the passage.) A list of decision rules was generated to guide the content analysis and summarization in these situations, and a 6-point rating system was developed to describe the level of specificity of the information sources. (See Appendice 10, 11 & 12) Appendix 13 contains sample items for each cell of the Math, Reading, and Writing matrices for which at least one state had test items.

An attempt was made to analyze all commercial, norm-referenced tests used by several of the states. Specimen Sets were ordered directly from the publisher. Unfortunately, not all commercial tests were received in time to be analyzed for this study. However, those included do provide a kind of sample of what such tests typically include.

After each state's materials had been examined, the data were summarized for each content area for 4 grade groupings: grades 1-3, 4-6, 7-9, 10-12. These summaries included the total number of test items and the number of different subskills tested in each major-skill-area-by-hierarchy-level cell in the matrix. For reading and writing, the number of cells for which test items occurred was relatively small (6 different cells). However, for math, the number of cells was larger, so the summary was done slightly differently. Numbers of items and subskills tested were summarized separately for each of 5 major skill areas and 4 hierarchical levels rather than the 20 different cells that would have resulted from crossing these axes. This method provided a relatively simple picture while still indicating the breadth and depth of content and cognitive level. In addition, a separate 20-cell math matrix of numbers of items and subskills was created for 5 major states at grades 4-6 and 7-9. Included on the summaries is each state's information source rating, which provided a measure of our confidence that what was reported is actually measured by the state's tests.

For the purpose of this study, "breadth" was viewed as the spread of test items across major skill areas and across the cognitive hierarchy within a given content area. The greater the number of different subskills, skill areas and hierarchy levels at which a state has test items, the greater the "breadth." "Depth" was defined as the number of test items for a given subskill at a given level of the hierarchy. The greater the number of items,the greater the "depth" for that particular subskill. As discussed earlier, other things being equal, broader tests with greater depth of coverage are considered to be "better".

In addition, lists of states were compiled for each of the criteria below:

1. states with "breadth" in any content area
2. states with "depth" in most of the skill areas of reading, math, or writing
3. states which emphasized higher order subskills
   (e.g. for reading: inferential & evaluative comprehension
   for math: any content requiring the 3rd or 4th level cognitive skill: explaining, translating, judging, or problem solving)
   for writing: organization & writing sample
4. states with items on attitude toward the content area
5. states with writing sample tests, by grade level
6. states which provided good documentation of their tests

Basic Results

The detailed content examination of state tests is provided in Appendices 15 - 18. (The key for interpreting these detailed summaries appears as Appendix 14.) These tables do depict the diversity of emphasis among the states in the material chosen for statewide testing. Some states sample broadly across skill areas with many subskills and many items per subskill (e.g., 250 items covering 30 subskills, typically matrix sampled); others measure many subskills with only a few items (e.g., 50 items covering 20 subskills); while still others test only a few subskill areas with lots of items (e.g., 80 items covering 8 subskills). Later in this chapter, we provide selected examples of what we view to be exemplary practice from the perspective of a broad-based, in-depth, balanced distribution of content with significant sampling of higher order skills.

For the present, however, we seek a simpler depiction of coverage for the purposes of selecting skills to concentrate on in an exploratory study. To accomplish this task, the content reported in Appendices 15 - 17 was used to develop state-by-skill area matrices for reading, math, and writing at each of the four grade level clusters. The entries in these matrices were coded as follows:

SKILL CODES
1 = State test includes at least one test item in the skill area
0 = State test does not include any items in the skill area
blank = No State test reported at this grade level
N = State tests at this grade but insufficient information on hand to determine what content was tested

The 12 state-by-skill area matrices were analyzed by Sato's Student Problem Chart procedure (See Harnisch (1983) for a description). This procedure (a) reordered the states vertically so that those testing in the most skill areas appear first and those testing in the fewest skill areas appear last, and (b) reordered the skill areas horizontally so that those skill areas tested most often by states appear first and those skill areas tested least often by states appear last. A summary table of number of states testing in a given skill area (as well as other information not reported here) was also generated.

The resulting matrices are reported in Tables 4.1-4.12. To visually simplify interpretation, a "." is used in place of a "1" when a state tested in the given skill area. Thus the meaning of the first row of data from Table 4.1 (Reading Grades 1-3) is that the state-developed minimum competency test in Alabama (first test listed for Alabama in Appendix 15) includes items from all five skill areas (word attack, vocabulary, literal comprehension, inferential comprehension, study skills). The same holds true for California, Hawaii, Kansas, Nevada, South Carolina, and Texas at these grade levels. Twenty eight states do not test in grades 1-3 and we have no information about Tennessee's test. None of the remaining states tested in all five skill areas according to the table.

Interpretation of skill area emphasis proceeds in a similar fashion. According to Table 4.1, items in the skill area of inferential comprehension were included in the most states (21) while study skills items (I) were included in the fewest (11). Note that a different skill ordering can occur at other grade intervals. For example, word attack skills were tested in the fewest states at grades 4-6 (and other grades for that matter).

One more feature of these tables deserves mention before proceeding with an examination of the results. The skill areas covered in some states are atypical for states testing in a given number of skill areas. For example, although inferential comprehension was the most popular skill area, Louisiana's test for grades 1-3 contains no items in this skill area but tests in all four remaining area. Florida's test apparently contains no literal comprehension items though the remaining skill areas are covered. When this type of analysis is applied to student test item responses, an atypical pattern is usually interpreted to reflect spotty student learning, guessing, or fundamental misunderstandings of certain concepts. In this present case, these atypical patterns could reflect a state's personalized curriculum emphasis, or perhaps simply the inadequacy of our classification efforts. We will try to note the occurence of such patterns as we consider the various tables.

Reading. We will consider each grade cluster separately, focussing on main trends and unique patterns of coverage. The discussion of grades 1-3 (Table 4.1) was basically provided in our examples. Only 22 states even test in this grade span (note Alabama has 2 testing programs); those that do tend to include items from every area except study skills. In addition to the atypical patterns of testing already mentioned in Florida and Louisiana, Arkansas's test does not include Vocabulary items but tests in the remaining areas.

There are 41 separate testing programs operating in the area of reading at grades 4-6 (Table 4.2); 3 states (Alabama, South Carolina,and Wisconsin) maintain 2 separate programs in this grade span. At least 18 programs test in all 5 skill areas while only 11 states do not test at all. A majority of states test in every skill area except word attack skills. The only apparent anomaly is again Arkansas's lack of coverage of vocabulary while testing in the remaining areas.

In grades 7-9 (Table 4.3), there are 42 separate test administrations (and 36 states testing) in reading. At least 20 programs test in all 5 skill areas while 12 states did not report testing at this grade span as of Fall 1985 (Subsequently, Indiana and South Dakota have started testing in grade 8.). Only word attack skills are tested in less than half the states while items on inferential and literal comprehension appear on at least

TABLE 4.1

STATE TESTING PROGRAMS READING CONTENT INDICATORS
ANALYSIS OF READING GRADES 1 - 3 DATE: JULY 1985

States	Skills Tested	Skill ILVWS
01AL1	5	.....
05CA	5	.....
11HI	5	.....
16KS	5	.....
28NV	5	.....
40SC	5	.....
43TX	5	.....
01AL2	4	0
03AZ	4	0
04AR	4	..0..
08DE	4	0
09FL	4	.0...
17KY	4	0
18LA	4	0....
20MD	4	0
31NM	4	0
33NC	4	0
38PA	4	...0.
48WV	4	0
14IN	3	...00
10GA	2	..000
32NY	1	.0000
02AK	0	
06CO	0	
07CT	0	
12ID	0	
13IL	0	
15IA	0	
19ME	0	
21MA	0	
22MI	0	
23MN	0	
24MS	0	
25MO	0	
26MT	0	
27NE	0	
29NH	0	
30NJ	0	
34ND	0	
35OH	0	
36OK	0	
37OR	0	
39RI	0	
41SD	0	
42TN	0	NNNNN
44UT	0	
45VT	0	
46VA	0	
47WA	0	
49WI	0	
50WY	0	

****** SKILLS STATISTICS ******

Permuted Skill Code	No. of States	Percent Testing
I	21	41.2
L	20	39.2
V	19	37.3
W	18	35.3
S	11	21.6

NOTES:

1) W = Word Attack
V = Vocabulary
L = Literal Comprehension
I = Inferential Comprehension
S = Study Skills

2) For states with more than 1 testing program at a given range of grade level multiple sets of codes are provided and tests are labled by numbers as well as state indicated (e.g., AL1, AL2).

3) For states for whom test content specifications were not available at the time of coding, the code N (no data) is reported in the table.

4) The number of states in a given skill area include all test versions from a state and excludes states for whom test specifications were not available at the time of coding.

BEST COPY AVAILABLE

TABLE 4.2

STATE TESTING PROGRAMS   READING CONTENT INDICATORS
ANALYSIS OF READING GRADES 4 - 6   DATE: JULY 1985

States	Skills Tested	Skill ILVSW
01AL1	5	.....
02AK	5	.....
05CA	5	.....
08DE	5	.....
11HI	5	.....
16KS	5	.....
17KY	5	.....
18LA	5	.....
22MI	5	.....
23MN	5	.....
26MT	5	.....
28NV	5	.....
29NH	5	.....
31NM	5	.....
37OR	5	.....
40SC1	5	.....
40SC2	5	.....
49WI2	5	.....
01AL2	4	0
03AZ	4	0
04AR1	4	..0..
04AR2	4	0
07CT1	4	0
09FL	4	...0.
20MD	4	0
33NC	4	0
38PA	4	0
43TX	4	0
44UT	4	0
46VA	4	0
47WA	4	0
48WV	4	...0.
10GA	3	...00
13IL	3	...00
14IN	3	...00
19ME	3	..0.0
25MO	3	..0.0
49WI1	3	..0.0
07CT2	2	..000
32NY	1	.0000
06CO	0	
12ID	0	
15IA	0	
21MA	0	
24MS	0	
27NE	0	
30NJ	0	
34ND	0	
35OH	0	
36OK	0	
39RI	0	NNNNN
41SD	0	
42TN	0	NNNNN
45VT	0	
50WY	0	NNNNN

**** SKILLS STATISTICS ****

Permuted Skill Code	No. of States	Percent Tested
I	40	72.7
L	39	70.9
V	34	61.8
S	33	60.0
W	21	38.2

NOTES:

1) W = Word Attack
V = Vocabulary
L = Literal Comprehension
I = Inferential Comprehension
S = Study Skills

2) For states with more than 1 testing program at a given range of grade level multiple sets of codes are provided and tests are labled by numbers as well as state indicated (e.g., AL1, AL2).

3) For states for whom test content specifications were not available at the time of coding, the code N (no data) is reported in the table.

4) The number of states in a given skill area include all test versions from a state and excludes states for whom test specifications were not available at the time of coding.

BEST COPY AVAILABLE

TABLE 4.3

STATE TESTING PROGRAMS   READING CONTENT INDICATORS
ANALYSIS OF READING GRADES 7 - 9   DATE: JULY 1985

States	Skills Tested	Skill ILVSW
01AL1	5	.....
02AK	5	.....
05CA	5	.....
08DE	5	.....
09FL	5	.....
16KS	5	.....
17KY	5	.....
18LA	5	.....
22MI	5	.....
23MN1	5	.....
23MN2	5	.....
29NH	5	.....
30NJ2	5	.....
31NM	5	.....
37OR	5	.....
40SC1	5	.....
40SC2	5	.....
42TN	5	.....
48WV	5	.....
49WI2	5	.....
01AL2	4	0
03AZ	4	0
07CT	4	0
12ID	4	0
20MD2	4	0
30NJ1	4	0
33NC	4	0
38PA	4	0
47WA	0	0
43TX	4	0
04AR	3	...00
10GA	3	...00
13IL	3	...00
14IN	3	...00
19ME	3	..0.0
20MD1	3	..0.0
28NV	3	..0.0
49WI1	3	..0.0
32NY	1	.0000
06CO	0	
11HI	0	NNNNN
15IA	0	
21MA	0	
24MS	0	
25MO	0	
26MT	0	NNNNN
27NE	0	
34ND	0	
35OH	0	
36OK	0	
39RI	0	NNNNN
42SD	0	
44UT	0	
34VT	0	
46VA	0	NNNNN

**** SKILLS STATISTICS ****

Permuted Skill Code	No. of States	Percent Tested
I	38	69.1
L	37	67.3
V	33	60.0
S	33	60.0
W	20	36.4

NOTES:

1) W = Word Attack
V = Vocabulary
L = Literal Comprehension
I = Inferential Comprehension
S = Study Skills

2) For states with more than 1 testing program at a given range of grade level multiple sets of codes are provided and tests are labled by numbers as well as state indicated (e.g., AL1, AL2).

3) For states for whom test content specifications were not available at the time of coding, the code N (no data) is reported in the table.

4) The number of states in a given skill area include all test versions from a state and excludes states for whom test specifications were not available at the time of coding.

TABLE 4.4

STATE TESTING PROGRAMS   READING CONTENT INDICATORS
ANALYSIS OF READING GRADES 10 - 12   DATE: JULY 1985

States	Skills Tested	Skill ILSVW
01AL.	5	.....
07CT	5	.....
18LA	5	.....
22MI	5	.....
23MN	5	.....
29NH	5	.....
30NJ2	5	.....
40SC	5	.....
42TN	5	.....
05CA	4	0
09FL1	4	..0..
11HI	4	0
16KS	4	0
26MT	4	0
30NJ1	4	0
33NC	4	0
37OR	4	0
44UT	4	0
09FL2	3	...00
10GA	3	...00
13IL	3	..0.0
19ME	3	...00
28NV	3	...00
38PA	3	..0.0
49WI	3	...00
25MO	1	.0000
32NY	1	.0000
02AK	0	
03AZ	0	NNNNN
04AR	0	NNNNN
06CO	0	
08DE	0	NNNNN
12ID	0	
14IN	0	
15IA	0	
17KY	0	NNNNN
20MD	0	
21MA	0	
24MS	0	
27NE	0	
31NM	0	
34ND	0	
35OH	0	
36OK	0	
39RI	0	NNNNN
41SD	0	
43TX	0	
45VT	0	
46VA	0	NNNNN
47WA	0	NNNNN
48WV	0	NNNNN
50WY	0	NNNNN

**** SKILLS STATISTICS ****

Permuted Skill Code	No. of States	Percent Tested
I	27	51.9
L	25	48.1
S	22	42.3
V	20	38.5
W	10	19.2

NOTES:

1) W = Word Attack
V = Vocabulary
L = Literal Comprehension
I = Inferential Comprehension
S = Study Skills

2) For states with more than 1 testing program at a given range of grade level multiple sets of codes are provided and tests are labled by numbers as well as state indicated (e.g., AL1, AL2).

3) For states for whom test content specifications were not available at the time of coding, the code N (no data) is reported in the table.

4) The number of states in a given skill area include all test versions from a state and excludes states for whom test specifications were not available at the time of coding.

BEST COPY AVAILABLE

37 tests. The patterns of content coverage are highly consistent across all states testing during this grade span.

Fewer testing programs are operated at grades 10-12 than at grades 4-6 and 7-9 (Table 4.4). There are 36 programs operating in 35 states, according to our data (Note that we failed to receive specimen sets from several commercial tests at this grade span.). Inferential and literal comprehension are still the most popular testing areas while coverage of vocabulary has dropped and word attack skills virtually disappeared. Again, patterns of content coverage are relatively uniform (some states test in vocubulary but not study skills).

If only one reading skill area and grade level were to be included in the exploratory study, the choice apparently boils down to either inferential or literal comprehension at either grades 4-6 or grades 7-9. An examination of the detailed summaries in Appendix 15 (and our study files) for these grade spans suggest that literal comprehension is likely to be a better skill area for the study. The basis for this judgment is some indication of greater uniformity across states in subskills tested in literal comprehension. When we examined our earlier descriptions of grades tested and dates of test administration more carefully, it appeared that there was more uniformity of practice in the older grade span where spring testing in grade 8 predominates. We return to this discussion of target grades and content areas later.

Mathematics. There are only 25 separate testing programs in 22 states in mathematics at grades 1-3 (Table 4.5). Most states operating a testing program test in the skill areas of numbers and numeration and measurement. According to our data, New York has a somewhat unusual topic coverage, skipping measurement and geometry but testing in statistics (the only state to do so at this grade span).

In grades 4-6 (Table 4.6), 39 testing programs in mathematics are administered by 36 states. At least 9 states test in all five skill areas and at least half the states test every area except statistics. Numbers and numeration and measurement are most frequently tested. Again, New York's apparent interest in statistics and lack of interest in measurement is the only atypical pattern.

Forty two (42) testing programs in mathematics are administered by 36 states in grades 7-9 (Table 4.7). At least 34 states test in 4 skill areas with numbers and numeration, measurement, and geometry the most popular. New York still avoids measurement at this grade span while Florida does not test in the geometry area.

Just as in reading, the number of testing programs drops rapidly in mathematics at grades 10-12 (Table 4.8). Eighteen states do not administer a mathematics test at this grade span. Numbers and numeration is still the most popular skill area, but

TABLE 4.5

STATE TESTING PROGRAMS   MATH CONTENT INDICATORS
ANALYSIS OF MATH GRADES 1 - 3   DATE: JULY 1985

State	Skills Tested	Skill NMVGS
01AL1	4	0
01AL2	4	0
03AZ	4	0
05CA	4	0
10GA	4	0
11HI	4	0
14IN	4	0
16KS	4	0
18LA2	4	0
20MD	4	0
23MN2	4	0
28NV	4	0
33NC	4	0
38PA	4	0
04AR	3	...00
08DE	3	...00
17KY	3	...00
23MN1	3	...00
32NY	3	.0.0.
09FL	2	..000
18LA1	2	..000
31NM	2	..000
43TX	2	.0.00
48WV	2	..000
02AK	0	
06CO	0	
07CT	0	
11HI2	0	
12ID	0	
13IL	0	
15IA	0	
19ME	0	
21MA	0	
22MI	0	
24MS	0	
25MO	0	
26MT	0	
27NE	0	
29NH	0	
30NJ	0	
34ND	0	
35OH	0	
36OK	0	
37OR	0	
39RI	0	
40SC	0	
41SD	0	
42TN	0	NNNNN
44UT	0	
45VT	0	
46VA	0	
47WA	0	
49WI	0	
50WY	0	NNNNN

**** SKILLS STATISTICS ****

Permuted Skill Code	No. of States	Percent Testing
N	23	43.4
M	21	39.6
V	19	35.8
G	13	24.5
S	1	1.9

NOTES:

1) N = #s & Numeration
V = Variables
G = Geometry
M = Measure
S = Statistics

2) For states with more than 1 testing program at a given range of grade level multiple sets of codes are provided and tests are labled by numbers as well as state indicated (e.g., AL1, AL2).

3) For states for whom test content specifications were not available at the time of coding, the code N (no data) is reported in the table.

4) The number of states in a given skill area include all test versions from a state and excludes states for whom test specifications were not available at the time of coding.

BEST COPY AVAILABLE

TABLE 4.6

STATE TESTING PROGRAMS MATH CONTENT INDICATORS
ANALYSIS OF MATH GRADES 4 - 6 DATE: JULY 1985

State	Skills Tested	Skill NMGVS
01AL1	5	.....
05CA	5	.....
10GA	5	.....
11HI	5	.....
14IN	5	.....
16KS	5	.....
25MO	5	.....
28NV	5	.....
38PA1	5	.....
01AL2	4	0
03AZ	4	0
07CT	4	0
08DE	4	0
13IL	4	0
17KY	4	0
18LA	4	0
20MD	4	0
23MN	4	0
26MT	4	0
31NM	4	0
33NC	4	0
38PA2	4	0
40SC	4	0
43TX	4	0
44UT	4	0
47WA	4	0
48WV	4	0
49WI	4	0
02AK	3	...00
04AR1	3	...00
04AR2	3	...00
22MI	3	...00
32NY	3	.0.0.
46VA	3	...00
09FL	2	..000
29NH	2	..000
37OR	2	..000
06CO	0	
12ID	0	
15IA	0	
19ME	0	NNNNN
21MA	0	
24MS	0	
27NE	0	
30NJ	0	
34ND	0	
35OH	0	
36OK	0	
39RI	0	NNNNN
41SD	0	
42TN	0	NNNNN
45VT	0	
50WY	0	

**** SKILLS STATISTICS ****

Permuted Skill Code	No. of States	Percent Testing
N	36	67.9
M	35	66.0
G	33	62.3
V	27	50.9
S	9	17.0

NOTES:

1) N = #s & Numeration
V = Variables
G = Geometry
M = Measure
S = Statistics

2) For states with more than 1 testing program at a given range of grade level multiple sets of codes are provided and tests are labled by numbers as well as state indicated (e.g., AL1, AL2).

3) For states for whom test content specifications were not available at the time of coding, the code N (no data) is reported in the table.

4) The number of states in a given skill area include all test versions from a state and excludes states for whom test specifications were not available at the time of coding.

BEST COPY AVAILABLE

TABLE 4.7

STATE TESTING PROGRAMS   MATH CONTENT INDICATORS
ANALYSIS OF MATH GRADES 7 - 9   DATE: JULY 1985

State	Skill Tested	Skill NGMVS
01AL1	5	.....
05CA	5	.....
07CT1	5	.....
07CT2	5	.....
07CT3	5	.....
10GA	5	.....
18LA	5	.....
20MD1	5	.....
29NH	5	.....
30NJ2	5	.....
38PA1	5	.....
38PA2	5	.....
42TN	5	.....
43TX	5	.....
01AL2	4	0
02AK	4	0
03AZ	4	0
04AR	4	...0.
08DE	4	0
12ID	4	...0.
13IL	4	0
16KS	4	...0.
17KY	4	0
20MD2	4	0
22MI	4	0
23MN1	4	0
23MN2	4	0
28NV	4	0
30NJ1	4	0
31NM	4	0
32NY	4	..0..
33NC	4	0
40SC	4	0
47WA	4	0
49WI1	4	0
09FL	3	.0..0
37OR	1	.0000
06CO	0	
11HI	0	NNNNN
14IN	0	
15IA	0	
19ME	0	NNNNN
21MA	0	
24MS	0	
25MO	0	
26MT	0	NNNNN
27NE	0	
34ND	0	
35OH	0	
36OK	0	
39RI	0	NNNNN
41SD	0	
44UT	0	
45VT	0	
46VA	0	NNNNN
50WY	0	NNNNN

**** SKILLS STATISTICS ****

Permuted Skill Code	No. of States	Percent Tested
N	36	64.3
G	34	60.7
M	34	60.7
V	32	57.1
S	18	32.1

NOTES:

1) N = #s & Numeration
V = Variables
G = Geometry
M = Measure
S = Statistics

2) For states with more than 1 testing program at a given range of grade level multiple sets of codes are provided and tests are labled by numbers as well as state indicated (e.g., AL1, AL2).

3) For states for whom test content specifications were not available at the time of coding, the code N (no data) is reported in the table.

4) The number of states in a given skill area include all test versions from a state and excludes states for whom test specifications were not available at the time of coding.

BEST COPY AVAILABLE

TABLE 4.8

STATE TESTING PROGRAMS   MATH CONTENT INDICATORS
ANALYSIS OF MATH GRADES 10 - 12   DATE: JULY 1985

State	Skills Tested	Skill NVGMS
05CA	5	.....
07CT	5	.....
09FL1	5	.....
10GA	5	.....
16KS	5	.....
22MI	5	.....
23MN	5	.....
25MO	5	.....
29NH	5	.....
33NC	5	.....
38PA	5	.....
42TN	5	.....
01AL	4	0
13IL	4	0
18LA	4	0
26MT	4	0
28NV	4	0
32NY	4	...0.
44UT	4	0
11HI	2	.00.0
09FL2	1	.0000
37OR	1	.0000
02AK	0	
03AZ	0	NNNNN
04AR	0	NNNNN
06CO	0	
08DE	0	NNNNN
12ID	0	
14IN	0	
15IA	0	
17KY	0	NNNNN
19ME	0	
20MD	0	
21MA	0	
24MS	0	
27NE	0	
30NJ	0	
31NM	0	
34ND	0	
35OH	0	
36OK	0	
39RI	0	NNNNN
40SC	0	NNNNN
41SD	0	
43TX	0	
45VT	0	
46VA	0	NNNNN
47WA	0	NNNNN
48WV	0	NNNNN
49WI	0	NNNNN
50WY	0	NNNNN

**** SKILLS STATISTICS ****

Permuted Skill Code	No. of States	Percent Testing
N	22	43.1
V	19	37.3
G	19	37.3
M	19	37.3
S	13	25.5

NOTES:

1) N = #s & Numeration
V = Variables
G = Geometry
M = Measure
S = Statistics

2) For states with more than 1 testing program at a given range of grade level multiple sets of codes are provided and tests are labled by numbers as well as state indicated (e.g., AL1, AL2).

3) For states for whom test content specifications were not available at the time of coding, the code N (no data) is reported in the table.

4) The number of states in a given skill area include all test versions from a state and excludes states for whom test specifications were not available at the time of coding.

BEST COPY AVAILABLE

the differences in emphasis among variables, geometry and measurement has disappeared. New York still excludes measurement while Hawaii includes it but excludes variables and geometry.

As in reading the choices for the exploratory study are between two skill areas (numbers and numeration or measurement) at two grade spans (4-6 or 7-9). An examination of the detailed summaries of content coverage does not provide much guidance in choosing between the two topics although New York would be excluded if measurement were the chosen area. The choice among grade spans must again rely on a more detailed examination of testing conditions as there are 36 states administering testing programs in either grade span. Spring testing in grade 8 occurs most frequently here as it does in reading.

Writing. We will devote less time to the discussion of writing because testing in this area is less widespread than in mathematics or reading and the Panelists expressed less interest in this area for that reason. Moreover, a note of caution is warranted about overinterpreting the results on the prevalence of writing at the various grade levels. Virtually all of the content classified as writing comes from indirect writing assessments rather than from writing samples. In fact much of this content is what might also be called language arts (conventions or grammar).

Despite the increased emphasis in recent years in direct writing assessments, the pattern of testing in this area is still quite poor (Tables 4.9-4.12). Only in the areas of conventions, word usage, and grammar do as many as half the states test and even then only in the grade spans 4-6 and 7-9. Only 3 or 4 states include items in all five skill areas at any given grade span. The collection of writing samples occurs infrequently even at the higher grades with the roughly 15 states collecting this data at grades 7-9 representing the largest sample of participating states. With the renewed interest in critical thinking coming on top of the interest in direct writing assessment, this area of testing should continue to grow and change in the coming years.

Exemplary Practices

Before proceeding to the recommendations regarding skill areas and grades proposed for the exploratory study, we want to briefly highlight exemplary practices that emerged in our examination. Three different aspects of practice will be emphasized: spread of items across subskills, depth of coverage within subskills, and significant coverage of higher order skills.

Significant numbers of states spread test items across a wide range of skill areas in at least one content area. The breadth of coverage was greatest in reading; 11 separate states were identified that exhibited broad coverage for at least one grade span. Alabama, California, Kansas, Florida, Louisiana, Michigan, Minnesota, New Hampshire, South Carolina, and Tennessee, had the most instances of tests with broad coverage.

TABLE 4.9

STATE TESTING PROGRAMS   WRITING CONTENT INDICATORS
ANALYSIS OF WRITING GRADES 1 - 3   DATE: JULY 1985

States	Skills Tested	Skill CWGOS
05CA	4	0
20MD	4	0
43TX	4	...0.
01AL1	3	..0.0
01AL2	3	...00
03AZ	3	...00
08DE	3	...00
11HI1	3	...00
14IN	3	.0.0.
17KY	3	...00
18LA	3	..00.
28NV	3	...00
31NM	3	...00
33NC	3	...00
48WV	3	...00
09FL	2	.00.0
02AK	0	
04AR	0	
06CO	0	
07CT	0	
10GA	0	
11HI2	0	NNNNN
12ID	0	
13IL	0	
15IA	0	
16KS	0	
19ME	0	
21MA	0	
22MI	0	
23MN	0	
24MS	0	
25MO	0	
26MT	0	
27NE	0	
29NH	0	
30NJ	0	
32NY	0	
34ND	0	
35OH	0	
36OK	0	
37OR	0	
38PA	0	
39RI	0	
40SC	0	
41SD	0	
42TN	0	
44UT	0	
45VT	0	
46VA	0	
47WA	0	
49WI	0	
50WY	0	

**** SKILLS STATISTICS ****

Permuted Skill Code	No. of States	Percent Tested
C	16	30.8
N	14	26.9
G	13	25.0
O	4	7.7
S	3	5.8

NOTES:

1) C = Conventions (e.g., spell, capit., punct.)
G = Grammar (sentence structure)
W - Word Usage
O = Organization
S = Writing Sample

2) For states with more than 1 testing program at a given range of grade level multiple sets of codes are provided and tests are labled by numbers as well as state indicated (e.g., AL1, AL2).

3) For states for whom test content specifications were not available at the time of coding, the code N (no data) is reported in the table.

4) The number of states in a given skill area include all test versions from a state and excludes states for whom test specifications were not available at the time of coding.

BEST COPY AVAILABLE

TABLE 4.10

STATE TESTING PROGRAMS WRITING CONTENT INDICATORS
ANALYSIS OF WRITING GRADES 4 - 6 DATE: JULY 1985

States	Skills Tested	Skill CWGOS
37OR	5	.....
01AL2	4	0
03AZ	4	0
05CA	4	0
07CT1	4	..0..
08DE	4	0
17KY	4	0
20MD	4	0
31NM	4	0
33NC	4	0
38PA	4	0
40SC1	4	0
43TX	4	...0.
47WA	4	0
48WV	4	0
49WI	4	0
01AL1	3	..0.0
04AR	3	...00
07CT2	3	..00.
09FL	3	.0..0
11HI	3	...00
13IL	3	...00
14IN	3	.0.0.
28NV	3	...00
44UT	3	..0.0
46VA	3	...00
25MO	2	.00.0
19ME	1	0000.
26MT	1	.0000
32NY	1	0000.
02AK	0	
06CO	0	
10GA	0	
12ID	0	
15IA	0	
16KS	0	
18LA	0	
21MA	0	
22MI	0	NNNNN
23MN	0	NNNNN
24MS	0	
27NE	0	
29NH	0	
30NJ	0	
34ND	0	
35OH	0	
36OK	0	
39RI	0	
40SC2	0	NNNNN
41SD	0	
42TN	0	
45VT	0	
50WY	0	NNNNN

**** SKILLS STATISTICS ****

Permuted Skill Code	No. of States	Percent Tested
C	28	52.8
W	24	45.3
G	22	41.5
O	19	35.8
S	7	13.2

NOTES:

1) C = Conventions (e.g., spell, capit., punct.)
G = Grammar (sentence structure)
W = Word Usage
O = Organization
S = Writing Sample

2) For states with more than 1 testing program at a given range of grade level multiple sets of codes are provided and tests are labled by numbers as well as state indicated (e.g., AL1, AL2).

3) For states for whom test content specifications were not available at the time of coding, the code N (no data) is reported in the table.

4) The number of states in a given skill area include all test versions from a state and excludes states for whom test specifications were not available at the time of coding.

BEST COPY AVAILABLE

TABLE 4.11

STATE TESTING PROGRAMS WRITING CONTENT INDICATORS
ANALYSIS OF WRITING GRADES 7 - 9 DATE: JULY 1985

State	Skills Tested	Skill CGWOS
07CT1	5	.....
07CT3	5	.....
30NJ	5	.....
37OR	5	.....
01AL2	4	0
03AZ	4	0
05CA	4	0
07CT2	4	0
08DE	4	0
09FL	4	0
13IL	4	0
17KY	4	0
18LA	4	...0.
20MD2	4	0
31NM	4	0
33NC	4	0
38PA	4	0
40SC2	4	0
42TN	4	0
43TX	4	...0.
48WV	4	0
49WI	4	0
01AL1	3	.0..0
14IN	3	..00.
12ID	2	.000.
19ME	1	0000.
20MD1	1	0000.
28NV	1	0000.
32NY	1	0000.
02AK	0	
04AR	0	
06CO	0	
10GA	0	
11HI	0	
15IA	0	
16KS	0	
21MA	0	
22MI	0	NNNNN
23MN	0	NNNNN
24MS	0	
25MO	0	
26MT	0	
27NE	0	
29NH	0	
34ND	0	
35OH	0	
36OK	0	
39RI	0	
40SC1	0	NNNNN
41SD	0	
44UT	0	
45VT	0	
46VA	0	
47WA	0	
50WY	0	NNNNN

**** SKILLS STATISTICS ****

Permuted Skill Code	No. of States	Percent Tested
C	25	45.5
G	23	41.8
W	23	41.8
O	21	38.2
S	12	21.8

NOTES:

1) C = Conventions (e.g., spell, capit., punct.)
G = Grammar (sentence structure)
W = Word Usage
O = Organization
S = Writing Sample

2) For states with more than 1 testing program at a given range of grade level multiple sets of codes are provided and tests are labled by numbers as well as state indicated (e.g., AL1, AL2).

3) For states for whom test content specifications were not available at the time of coding, the code N (no data) is reported in the table.

4) The number of states in a given skill area include all test versions from a state and excludes states for whom test specifications were not available at the time of coding.

TABLE 4.12

STATE TESTING PROGRAMS   WRITING CONTENT INDICATORS
ANALYSIS OF WRITING GRADES 10-12   DATE: JULY 1985

States	Skills Tested	Skill CWOGS
07CT	5	.....
18LA	5	.....
30NJ	5	.....
09FL	4	0
13IL	4	0
25MO	4	0
38PA	4	0
42TN	4	0
01AL.	3	...00
05CA	3	.0..0
26MT	3	..0.0
44UT	3	...00
11HI	1	0000.
19ME	1	0000.
28NV	1	0000.
32NY	1	0000.
37OR	1	0000.
02AK	0	
03AZ	0	
04AR	0	
06CO	0	
08DE	0	
10GA	0	
12ID	0	
14IN	0	
15IA	0	
16KS	0	

States	Skills Tested	Skill CWGOS
17KY	0	
20MD	0	
21MA	0	
22MI	0	NNNNN
23MN	0	NNNNN
24MS	0	
27NE	0	
29NH	0	
31NM	0	
33NC	0	
34ND	0	
35OH	0	
36OK	0	
39RI	0	
40SC	0	NNNNN
41SD	0	
43TX	0	
45VT	0	
46VA	0	
47WA	0	
48WV	0	NNNNN
49WI	0	NNNNN
50WY	0	NNNNN

**** SKILLS STATISTICS ****

Permuted Skill Code	No. of States	Percent Tested
C	12	24.0
W	11	22.0
O	11	22.0
G	10	20.0
S	8	16.0

NOTES:

1) C = Conventions (e.g., spell, capit., punct.)
G = Grammar (sentence structure)
W = Word Usage
O = Organization
S = Writing Sample

2) For states with more than 1 testing program at a given range of grade level multiple sets of codes are provided and tests are labled by numbers as well as state indicated (e.g., AL1, AL2).

3) For states for whom test content specifications were not available at the time of coding, the code N (no data) is reported in the table.

4) The number of states in a given skill area include all test versions from a state and excludes states for whom test specifications were not available at the time of coding.

BEST COPY AVAILABLE

The number of states exhibiting depth of coverage (lots of items per subskill) in more than one content area were very few. California has deep coverage everywhere by our criteria while Alabama and Minnesota exhibited deep coverage in reading and math (Connecticut may have also but we did not complete the coding of its reading assessment). Most of the states who had broad coverage also managed to include a lot of test items for at least one skill area.

The testing of higher order skills is perhaps of greatest interest. At least 14 states included significant numbers of higher order skill items on their tests. California, Connecticut, Illinois, Kansas, Michigan, New York, Alabama, Oregon, Pennsylvania, and Indiana (new test) appear to stand out in this area.

Several states appear to have strong tests across the board. States with extensive, long-standing internally developed tests (e.g., California, Connecticut, Florida, Michigan, Kansas, Minnesota, Pennsylvania, and Illinois) tend to fare best according to our criteria. But there were several surprises. The positive showing of programs in Alabama, Louisiana, and South Carolina suggests that region of the country is not a determining factor in testing program quality. New York's well-respected testing programs do not compare favorably by our criteria but this could simply be lack of information on our part.

One other point is worth noting. Generally states who emphasize commercially available standardized tests do not fare well by the criteria we have used to characterize exemplary practices. Their performance may simply be underrated because we lacked test copies at the higher grades for most standardized tests. Or it could be an indication of these tests' conservative content strategy when compared with the presumably more locally sensitive tests developed directly by states.

Despite the somewhat rosy picture for testing of higher order skills in some states, most states have too little coverage in these skill areas to mount a broad based exploratory study. This is unfortunate if well-developed higher order skills are indeed the focus of the new curriculum reforms as it will be difficult to monitor the effects of reforms on these skills without more extensive test coverage at higher levels.

Summary and Recommendations

Our discussions in this chapter barely scratch the surface of the details of content of existing state tests and of tests just over the horizon in many states. Yet we have conducted by far the most extensive examination of the content of state tests to date. (Subsequent to the completion of our data collection, the Office of Technology Assessment contracted with Northwest Regional Laboratory to carry out yet another survey of state testing programs with an even more detailed focus on content coverage and changes in coverage over time. The results of that study are not yet available.)

While we were unable to carry out work to the point of developing an explicit set of indicators of content coverage, we were able to hone in on target areas and grades for the proposed exploratory study. After a careful examination of the test content data, information about grades tested and dates of test administration, the best candidates for the study appear to be the areas of literal comprehension and either numbers and numeration or measurement in mathematics at grades 7-9. The basic reasons for the content choices have already been provided (primarily frequency of testing at the target grades). The decision to focus on the same grade span in both content areas is an attempt to reduce complexity and costs and disrupt as few schools in a given state as possible.

The choice of grades 7-9 over grades 4-6 is based primarily on the number of deviations within the grade span from the single most frequent grade/test administration date combination. The grade level most frequently testing is grade 8 while the states testing in the grade 4-6 range are more evenly spread across grades.

Table 4.16 summarizes the testing conditions of States in grades 7-9 as of the Spring 1985. Of the 40 states who test at grades 7-9 (or planning to do so soon), 25 administer their tests in the spring to students in the 8th grade. This leaves only 15 states that currently test in this grade span who would have to either change their grades for testing, change their time of year, or do both. The other alternative for these states is to carry out the special studies of testing conditions to estimate the adjustments necessary to align their performance with that of spring testing of 8th graders. There are only three states (Michigan, Nevada, and West Virginia) in which both grade and date of testing do not match the target testing conditions.

The set of states who currently test in the proposed skill areas during grades 7-9 are depicted in Figure 4.1. Note that New York would be eliminated due to its idiosyncratic content coverage at this grade span. Without any modifications of current practices, comparisons would be workable in the South, the Far West, the East, and the Upper Midwest. As programs in states just starting their own assessment begin to develop, the picture will be even better. For instance, the states of Wyoming, Indiana, and South Dakota are just starting to collect testing data and Mississippi is due to begin by 1987. The trend is clearly in the direction of more testing and greater conformity in testing practices.

TABLE 4.13

7/24/85

READING

States w/wide "spread" across subskills: (by grade level)

	1-3	4-6	7-9	10-12
These states have 2 or more subskils in every skill area	AL	AL	AL	AL
	CA	OR	MN	LA
			OR	TN
			TN	
These have 1 or more subskills in each skill area	FL	SC	AL	FL (2 tests combined)
	KS	AL	CA	MI
	SC	CA	FL	MN
	TX	KS	KS	NH
		LA	LA	SC
		MI	MI	TN
		MN	NH	
		MT	NJ (phasing out)	
		NH	SC	

States w/most "depth" - i.e. most items per subskill

*CA [e.g. grade 1-3, WA= 60/3, Voc = 30/2, LC = 73/3, IC = 77/4, SS = 30/2]
MN
MT
NY - only on infer word goes in blank (entire reading test is this format)

States w/emphasis on higher order subskills ("IC"): (lots of items and/or lots of subskills)

1-3		4-6		7-9		10-12	
IN	30/6	CA	78/16	CA	235/15	CA	50/5
NY	56/1	MI	27/8	IL	10/7	KS	21/7
CA	77/4	NY	77/1	IN	35/7	LA	24/6
SC	/8	IN	35/7	KS	15/5	MI	26/8
		SC	/8	MI	24/7	MO	12/6 = Whole Test
		CT	24/11	CT	/16	MT	20/6
						NJ	43/11
						NY	77/1
						PA	34/7

States w/items on attitude toward reading:

Michigan (15 items at 4-6, 7-9, 10-12)
Montana (15 items at 4-6, 10-12)
Connecticut (CAEP)

TABLE 4.14

7/24/85

MATH

States with wide spread across 5 skill areas (by gr. level):

1-3	4-6	7-9	10-12
No states included "statistics" at this grade level. AL have 4 or CA more items IN in each of KS other skill LA areas	AL have 3 or CA more items GA in each of IN 5 skill areas KS MO CT	AL have 4 or CA more items LA in each of MN 5 skill areas CT GA 1-3 items is NH lowest amt. NJ in any of 5 PA skill areas TX	AL 4 or more CA items in MN each of 5 skill aeras KS 1-3 items GA is lowest MI amt. in MO any of 5 NH skill areas PA TN CT

States with "depth" (most items per subskills):

CA - the most
AL
FL

ID usually
KS a lot
LA of items
MI in "#s & Numeration"
MN
CT

States with emphasis on higher order subskills (3* & 4* in following chart)
(lots of items and/or lots of subskills)

1-3			4-6			7-9			10-12		
CA	20/4	37/5	CA	71/7	70/4	AL	8/2	44/5	AL	6/2	23/3
			MT	---	26/5	CA	100/11	105/5	CA	---	55/6
			PA	13/4	17/3	FL	---	10/1	FL	---	40/2
			CT	88/6	16/1	IL	2/2	17/3	IL	3/1	16/5
						KS	---	9/2	KS	---	51/3
						NJ	1/1 3/1	19/5 31/4			
						OR	10/4	15/3	MN	[?]	29/3
						CT	36/6	20/4	MT	---	47/5
									OR	1/1	20/2
									CT	3/2	10/4

States with items on attitude toward math:

CONNECTICUT

Also - only CT had items on computers and calculators plus some items on computer literacy in its lang. arts section of CAEP test.

TABLE 4.15

7/24/85

WRITING

STATES WITH WRITING SAMPLES:

		Gr.	1 - 3	4 - 6	7 - 9		10 - 12	Scoring Method*
(new)	1	Idaho			X			?
	2	Indiana	X	X	X			H
	3	Louisiana	X		X		X	P
	4	Maine		X	X		X	H,P,A
	5	Nevada			X	---->	X	H
	6	New Jersey			X	---->	X	?
	7	New York		X	X		X	H
	8	Oregon		X	X		X	H
	9	Texas	X	X	X			?
	10	Maryland			X			?
	11	Connecticut		X	X		X	H,A

STATES WITH QUESTIONS ON ATTITUDES TOWARD WRITING

Illinois
Montana
Connecticut

STATES WITH "SPREAD" ACROSS WRITING CONTENT:

California (esp. 1-3, 4-6 and 7-9)
Connecticut
Florida
Illinois (esp. 7-9, 10-12)
New Jersey
Oregon
Pennsylvania (voluntary test)
Tennessee

STATES WITH "DEPTH"

1. California - has most items per area
2. Alabama - medium amount of items per area

HIGHER ORDER WRITING SKILLS OTHER THAN WRITING SAMPLE ("OR" & "SM" columns)

California (judge student writing on specifics)
Connecticut (take notes; ID missing info. on outline)
Illinois (editing in 8th & 10th grades)
Oregon, Alabama (fill out forms; letter format)
Pennsylvania (judge relevance gr. 5, 8 & 11)

*Scoring Method Key:

H = Holistic P = Primary Trait
A = Analytic (Diagnostic Checklist)
? = Not specified in documents

TABLE 4.16

State Testing Conditions in Reading and Mathematics
Grades 7-9 as of Spring 1985

States Testing in Grade 8 During Spring (Feb-May), (N=25)

Alabama (formerly CAT, Now SAT)
Alaska (every 2 years)
Arizona (CAT)
California
Delaware (CTBS)
Florida (every 2 years, MCT)
Georgia (ITBS)
Idaho (MCT)
Illinois
Indiana (beginning Feb 1985, MCT)
Kansas (MCT)
Kentucky (CTBS)
Missouri (MCT)
Montana
New Mexico (CTBS)
New York (MCT)
Pennsylvania
Rhode Island (ITBS)
South Carolina (MCT)
South Dakota (beginning April 1985)
Tennessee (formerly MAT)
Virginia (SRA)
Washington (CAT)
Wisconsin (CTBS)
Wyoming (NAEP)

States Testing in Grades 7 or 9 During Spring (Feb-May), (N=7)

Arkansas (7, SRA)
Hawaii (9, MCT)
Louisiana (7)
New Jersey (9, MCT)
North Carolina (9, CAT)
Oregon (7, every yr. 1985+)
South Carolina (7, CTBS)

States Testing in Grade 8 During Fall or Winter, (N=6)

Connecticut (CAEP)
Hawaii (SAT)
Maine
Maryland (CAT)
Minnesota
New Hampshire (MCT beg. 1985)

States Testing in Grades 7 or 9 During Fall or Winter (N=3)

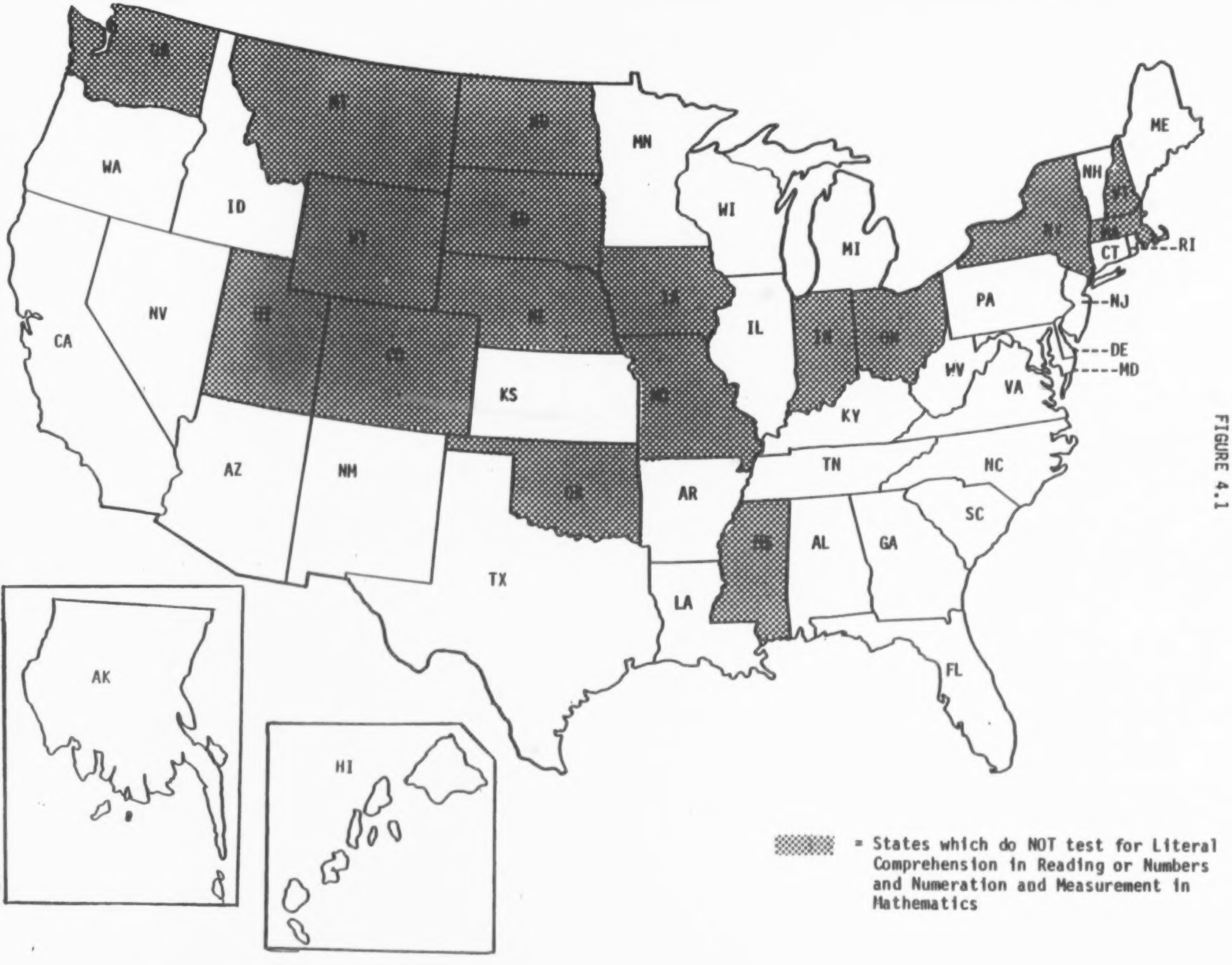
Michigan (7, MCT?)
Nevada (9, MCT)
West Virginia (9, CTBS)

No Grade 7 through 9 Testing (N=1)

Utah

No State Testing at any Grade (N=8)

Colorado
Iowa
Massachusetts
Nebraska
North Dakota
Ohio
Oklahoma
Vermont



OR
MT
ND
MN
ME
WA
ID
SD
WY
WI
NH
VT
NY
MA
MI
CT
RI
IA
PA
NJ
NV
CA
UT
NE
IL
IN
OH
CO
DE
MD
WV
KS
MO
VA
KY
AZ
NM
TN
NC
OK
AR
SC
MS
AL
GA
TX
LA
FL
AK
HI

▨ = States which do NOT test for Literal Comprehension in Reading or Numbers and Numeration and Measurement in Mathematics

FIGURE 4.1

Chapter 5
Examination of Reporting Practices and Auxilliary Information

Statement of the Problem

Within-state contrasts in achievement could be used to make between-state comparisons of performance. There are two types of within-state contrasts that could be of special interest:

1) Longitudinal Contrasts which examine trends in achievement test scores over time. There are two types of longitudinal contrasts that would be of interest:

a) Cohort repetitive trends, in which the same students are followed year-by-year, from grade-to-grade. For example, students are tested at Grade 1 in the first year, then followed over the years to grade 6. Some states do not track exactly the same students, but provide test information for all students at each successive grade level. Changes in cohort composition are confounded with instructional treatment when the data are not for the identical students at each point in time. When the data are for identical students, attrition may account for some of the observed trends.

b) Cohort replicative trends, in which successive groups of students at a given grade level are tested. For example, fourth graders are tested each year in reading. Trends over time will be confounded with changes in the student population at the grade level(s) tested.

2.) Subgroup Contrasts in which different groups within a state are contrasted to one another. Contrasting scores of students in different socio-economic status brackets, or contrasting the performance of different racial/ethnic groups are examples of contrasts within states that could form the basis of state-to-state comparisons. At a minimum, the definitions of the subgroups would have to be consistent across states in order to permit cross-state comparisons. Although states have federal models for some categories of classification (e.g., the Office for Civil Rights classification of race/ethnicity), they may not use these consistently in their achievement testing programs. In areas with lesser political salience, the definitions of subgroups could be quite varied.

Because longitudinal trends may be confounded with changes in cohort composition, the combination of subgroups and trend contrasts would provide basis for more accurate comparison. However, it is unlikely that many states will have information on the same subgroups (e.g., grade-level, racial/ethnic status, sex) tested in the same skill areas, over time. Even if such information were available, it is not likely to be reported in

---

J. Ward Keesling was primarily responsible for the preperation of this chapter.

the same metric across different states. For example, in our examination of reports from various states, statewide test performance was reported using the following metrics: grade equivalents, percent correct, percent scoring above a specified passing score, stanines, percentiles, and various standard scores. While scores reported in some of these metrics are often confused with each other, none are directly comparable. Moreover, states seldom report the necessary distributional information (e.g., standard deviations of performance for each year in a longitudinal series or for each subgroup in the case of subgroup contrasts) to permit transformation of reported scores to standardized units (gains in standard deviation units, subgroup contrasts expressed as effect sizes) that might be comparable across states.

A further problem with the mixture of metrics is that there is no absolute scale of comparison. If the data available are reduced to gains or subgroup contrast effects, there may be no way to recognize when one state is experiencing low gains or small subgroup contrasts due to ceiling effects, for example. However, even the simplest indicator (a + sign indicating gain vs. a - sign indicating loss) could serve, over time, as a signal that interesting differences were occurring. If blacks in one state show achievement gains from year-to-year over 4 to 5 years (3 to 4 differences) while blacks in a contiguous state show losses, no matter what the metric, there would be reason to examine the educational programs (and other factors) more thoroughly.

The problems with varying metrics are not restricted to the reporting of achievement. States gather certain types of auxiliary information using different scales. Definitions of school characteristics such as dropout rate, ADA, and type of community in which the school is located, and student characteristics such as parental education and occupation are not measured in a uniform manner even among the few states that collect them. Until a greater degree of uniformity of information collection is attained or some other means are developed to alleviate the metric problems with auxiliary variables, the use of state-collected auxiliary information as either additional indicators of context, resources, processes and outcomes or as a basis for subgroup classification for generating within-state performance contrasts will be severely limited.

Current Collection and Reporting Practices

Setting aside concerns about possible metric differences, the question remains whether extant state data can be used to generate within-state comparisons of the kinds discussed above.

During the telephone interviews, state testing program representatives were asked whether:

(a) they report longitudinal or time trend data and over what period if they did;

(b) they report achievement data for different subgroups of students, and how these were defined.

Copies of state reports were examined for evidence that they contained either trend information or subgroup results on achievement. The interviews and the examination of reports also produced data about the auxiliary information collected or reported as part of state testing programs.

Table 5.1 shows the combination of subgroup and auxiliary information that was detected in the interviews and/or in the examination of reports. It should be pointed out that most states used the subgrouping and auxiliary information to profile the composition of their student population; relationships between these characteristics and the achievement scores were not often explored. Some states collected this information but did not use it in their reports. This table may be an underestimation of the information available in raw form in the states because some data may be collected and not used in reports, and may also have been missed out in the interviews.

Table 5.2 is a more focussed examination of the state-by-state reporting of subgroup comparisons or longitudinal trends. It is also based upon the interviews and examinations of the reports we received. Tables 5.3 and 5.4 summarize the information in Table 5.2. Table 5.3 shows that 27 states in our sample of 36 had longitudinal data for a span of at least 3 years. Six states had no trend information, and two others had it, but did not report it.

Table 5.4 shows that about one third of the states in our sample of 36 report no information on subgroups. Sex and racial/ethnic background were the most frequently used subgroupings. Again, one or two states collect subgroup data but do not report it.

The next step in our examination of the state reports was to look at the specific nature of the longitudinal and subgroup contrasts that were reported to determine if they could form the basis of state-to-state comparisons. Because we could anticipate that race/ethnic background classifications might vary by state, it seemed prudent to focus on gender classification because it was frequently used and unlikely to vary by state. We chose to examine all states that had been cited as having both sex subgrouping and trend data of 3 years or more. This led us to examine more closely the reports of the following 13 states: Arizona, California, Connecticut, Louisiana, Maine, Minnesota, North Carolina, Pennsylvania, Rhode Island, South Carolina, Texas, Virginia, and Wisconsin. We focussed on the availability of achievement results for students in grades 7-9, in reading or math. This grade span was chosen because our analysis of the state testing programs had shown this to be a popular grade range in which to test (see Table 4.1 - 4.12). We looked for results on tests of literal comprehension in the reading area and on measurement or computational skills in the math area in order to
TABLE 1 - STATE TABLE

TABLE 5.1

revised 7/24/85

Auxiliary Information Collected or Reported By State Testing Program

States

Information	AL	AK	AZ	AR	CA	CO	CT	DE	FL	GA	HI	ID	IL	IN	IA	KS	KY	LA	ME	MD	MA	MI	MN	MS	MO	MT	NE	NV	NH	NJ	NM	NY	NC	ND	OH	OK	OR	PA	RI	SC	SD	TN	TX	UT	VT	VA	WA	WV	WI	WY
I. Students																																																		
A. Background																																																		
1. Chapter I				x	x																														x					x			x				x			
2. Chapter I - Migrant				x	x																														x								x				x			
3. Ethnic Group		x		x			x	x	x		x		x					x			x						x				x				x			x		x			x							
4. Free Lunch					x																																			x			x							
5. Grade Repeaters																																								x										
6. Language Status		x		x	x								x						x		x										x												x				x			
7. Years in Community					x																								x		x																			
8. Occupation (parent/s)					x																																													
9. Parent Educ.					x		x						x																						x															
10. RAP Programs																																															x			
11. Sex		x		x	x		x	x	x				x					x	x		x	x	x				x								x		x	x		x			x					x	x	
12. Special Ed.																x																			x					x			x				x			
13. Date of Birth					x		x																																											
14. Years in School					x																																													
15. Years in District					x																																													
16. Parent Support													x								x																													
17. Soc-Econ Status																		x												x																				
18. Family Size													x																																					
B. Curriculum Exposure - General																																																		
1. Curriculum Track																			x																															
2. Homework, Hours Spent					x		x																												x															
3. Instruction, Minutes of					x																																													

BEST COPY AVAILABLE

States cont. (2)

Information	AL	AK	AZ	AR	CA	CO	CT	DE	FL	GA	HI	ID	IL	IN	IA	KS	KY	LA	ME	MD	MA	MI	MN	MS	MO	MT	NE	NV	NH	NJ	NM	NY	NC	ND	OH	OK	OR	PA	RI	SC	SD	TN	TX	UT	VT	VA	WA	WV	WI	WY
I. Students																																																		
C. Student Attitudes & Activities - General																																																		
1. Attitude Toward Computers					x		x																																											
2. Attitude Toward School					x		x						x						x																											x				
3. Academic Self-Concept							x				x																																			x				
4. Educational Plans					x		x						x																																	x		x		
5. Career Plans					x																																									x		x		
6. Talk to Parents About School							x						x						x																															
7. Parental Encouragement							x						x						x																															
8. TV					x		x				x		x						x			x																												
9. Emotional Maturity							x																																							x				
10. Peer Relations																																														x				
11. Teacher Support													x																																					
12. Peer Support													x																																					
13. Attr. of Success													x																																					
14. School Climate					x		x						x																																					
15. Test Anxiety													x																																					
D. Student Attitudes & Activities - Reading																																																		
1. Read Newspaper							x						x						x																															
2. Read for Pleasure							x												x			x																												
3. Library Books for Non-School Assgnmt.							x												x																															
4. How Well Student Feels S/he's Been Taught Reading							x												x																															
5. Visit Reading Places													x						x																															

BEST COPY AVAILABLE

States cont. (3)

Information	AL	AK	AR	AZ	CA	CO	CT	DE	FL	GA	HI	ID	IL	IN	IA	KS	KY	LA	ME	MD	MA	MI	MN	MS	MO	MT	NE	NV	NH	NJ	NM	NY	NC	ND	OH	OK	OR	PA	RI	SC	SD	TN	TX	UT	VT	VA	WA	WV	WI	WY
I. Students																																																		
D. Student Attitudes & Activities - Reading																																																		
6. Request Extra Reading																						x																												
7. Talk About Reading							x												x																															
8. Completion of Specific English Courses					x		x																																											
9. Time on Homework in English					x		x																																											
10. Days of Homework in English					x																																													
11. Tests & Quizzes in Reading					x																																													
12. Hours/days Reading for Class Assignments					x		x																																											
13. Like Reading																																																x		
E. Student Attitudes & Activities - Writing																																																		
1. Write for Own Purposes							x						x																																					
2. Write Assignments in English Class					x		x						x																																					
3. Write Assignments in non-English Class													x																																					
4. How Often Write for School					x		x						x						x																															
5. Revise Writing							x						x						x																															
6. Teachers Talk with Students About Their Writing							x						x						x																															

States cont. (4)

Information	AL	AK	AZ	AR	CA	CO	CT	DE	FL	GA	HI	ID	IL	IN	IA	KS	KY	LA	ME	MD	MA	MI	MN	MS	MO	MT	NE	NV	NH	NJ	NM	NY	NC	ND	OH	OK	OR	PA	RI	SC	SD	TN	TX	UT	VT	VA	WA	WV	WI	WY
I. Students																																																		
E. Student Attitudes & Activities - Writing																																																	
7. How Well Student Feels S/he's Been Taught Writing													x						x																															
F. Student Attitudes & Activities - Math																																																	
1. Semesters of Math					x																																													
2. Completion of Specfic Math courses					x																																													
3. Time on Homework in Math					x																																													
4. Days of Homework in Math					x																																													
5. Tests & Quizzes in Math					x																																													
6. Like Math																																																x		
G. Other Specific Curriculum Activities																																																		
1. Completion of Specific Soc. Studies Courses					x																																													
2. Completion of Specific Foreign Language Courses					x																																													
3. Completion of Specific Art, Music, Drama Courses					x																																													
4. Completion of Specfic Science Courses					x																																													
5. Time on Homework in Soc. Studies					x																																													
6. Time on Homework in Science					x																																													

BEST COPY AVAILABLE

States cont. (5)

Information	AL	AK	AZ	AR	CA	CO	CT	DE	FL	GA	HI	ID	IL	IN	IA	KS	KY	LA	ME	MD	MA	MI	MN	MS	MO	MT	NE	NV	NH	NJ	NM	NY	NC	ND	OH	OK	OR	PA	RI	SC	SD	TN	TX	UT	VT	VA	WA	WV	WI	WY
I. Students																																																		
G. Other Specific Curriculum Activities																																																		
7. Days of Homework in Soc. Studies					x																																													
8. Days of Homework in Science					x																																													
9. Tests & Quizzes in Science					x																																													
10. Tests & Quizzes in Soc. Studies					x																																													
11. Like/Favorite Subj.																																																x		
II. Schools																																																		
A. Community Context																																																		
1. District Size		x					x																x																								x			
2. County/City; Region/District; City/Parish										x						x		x	x											x			x							x			x			x				
3. Urban/Suburban													x																	x																				
4. Community Type							x						x						x																															
5. District Loc.													x																																					
B. Socio-Economic Characteristics																																																		
1. AFDC					x																																													
2. Exceptionality																																	x																	
3. Migrant Child																																	x																	
4. District SES																														x																				
5. School Size/ADA										x			x			x				x																														
6. Mobility					x																																													
7. Free Lunch													x																																					
C. Staff & School Resources																																																		
1. Number of Professional Staff																				x																														

States cont. (6)

Information	AL	AK	AZ	AR	CA	CO	CT	DE	FL	GA	HI	ID	IL	IN	IA	KS	KY	LA	ME	MD	MA	MI	MN	MS	MO	MT	NE	NV	NH	NJ	NM	NY	NC	ND	OH	OK	OR	PA	RI	SC	SU	TN	TX	UT	VT	VA	WA	WV	WI	W
II. Schools																																																		
C. Staff & School Resources																																																		
2. Avg. Pupil/Staff																				x											x																			
3. Avg. Teacher Salary																				x																														
4. % Teachers With MA							x													x																														
5. Number Pupils Tested Per Region																																	x																	
6. Per Capita Income																																	x																	
7. Avg. Ed. Expenditures																																	x																	
8. Per Pupil Expenditures																																	x																	
9. Teacher Exper.							x																								x																			
10. Courses Offered by Curricular Field													x																																					
D. Other																																																		
1. Public/Private													x																			x								x										
2. Absence Rate																																	x																	
3. Class periods/ School day					x																																													
4. % Class time lost to Disruption & Distraction					x		x																																											
5. % of Teachers Pointing out Dangers of Drug use					x																																													
6. Drop out Rate								x																																										

BEST COPY AVAILABLE

TABLE 5.2

Assessment Report Contents

State	Subgroup Info	Longitudinal Info
ALABAMA	NO	NO
ALASKA	Race/Language	NO
ARIZONA	Race/Chap 1/Sex Language	4 years
CALIFORNIA	Sex/Language/ Parent Ed level/ Exposure to math	4 years
CONNECTICUT	Sex/Community/Urban	5 years
DELAWARE	NO	6 years (not reported)
FLORIDA	Available, but not reported (NO)	4 years
GEORGIA	Free Lunch/Region/ LEA enrollment	4 years
IDAHO	NO	NO
ILLINOIS	Language	2 years
KANSAS	District/Region/ School Enrollment	5 years
KENTUCKY	NO	3 yrs
LOUISIANA	Sex/Race/Soci-econ/City-parish	3 years
MAINE	Sex/Type of prog./ Language/Race/ [grade: on communitems] Region/Community type	3-5 yrs. (not yet reported)
MARYLAND	NO	NO
MICHIGAN	Sex	2 years
MINNESOTA	Sex	3 years
MISSOURI	NO	Yes/not reported

State	Subgroup Info	Longitudinal Info
MONTANA	NO	NO
NEVADA	Not reported	5 years
NEW HAMPSHIRE	NO	NO
NEW JERSEY	Urbanism/Classe	7 years
NEW MEXICO	Ethnicity/Language/ Yrs of residence	3 years
NEW YORK	Public vs. priv./ Community type	5 years
NORTH CAROLINA	Sex/Ethnicity/ Handicap/Homework/ Region/Chapter 1/ Parent educ.	4-6 years
OREGON	NO	4 yr/ 2 points
PENNSYLVANIA	Race/Sex/District	3 years
RHODE ISLAND	Sex/SES	4 years
SOUTH CAROLINA	Sex/Race/Chap 1/ Free lunch/Repeater/ Handicap/Gifted/District	4-6
TENNESSEE	NO	NO
TEXAS	Race/Sex/SES/Spec ed/Program/Language/ Region	4 years
UTAH	Student demography/ School sampling/strata	6 yrs/3 pts.,
VIRGINIA	NO Race/Sex/Handicap/ District	3 years 6 years
WASHINGTON	Race/Chap 1/Spec program/District	3-5 yrs/ Not reported
WEST VIRGINIA	NO	Not reported
WISCONSIN	Sex/Attitudes toward subjects	2-8 yrs

TABLE 5.3

State Reports of Longitudinal Trends

	Span of Years Reported On							
	8	7	6	5	4	3	2	No Report
Number of States	1	2	5	6	7	6	1	8
Cumulative Number of States	1	3	8	14	21	27	28	36

Notes:

1. 27 have at least 3 years of data they have reported on
2. One or two don't report trends every year, even if they test annually - therefore time points may not be the same in number for all these LEAs in the same category.

TABLE 5.4
State Reports of Subgroup Information

Subgroup Typology	Number of States Reporting
None	13
Sex	14
Race/Ethnic background	11
Region	7
Language Proficiency	7
Socio-Economic Status	5
Community Type (e.g., urban vs. rural)	5
Chapter 1 participant	4
District enrollment	4
Handicap	4
Type of School Program (may include chap 1 or handicap)	3
Parent Education	2

Reported by only one state each:

School enrollment
Exposure to instruction
Years of resdence
Public vs. Private school
Student demography
Homework
Gifted
Repeating a grade
Attitudes toward subject matter

TABLE NOTES:

1. Based on 36 states with interviews or analysed reports.
2. Category schemes with the same name may be different from state-to-state.

make the test content as comparable as possible. When we were unable to find test results on these subskills, we reported the results for TOTAL math or TOTAL reading instead.

Despite our attempts to homogenize content, these can still be considerable variations so comparisons can only be crude at best. A brief synopsis of our findings for each state follows:

ARIZONA:

Uses CAT tests in the Spring. Metric: percentile. Grade 8

Sex contrast: 1984

	Male	Female
Reading	60	61
Math	62	65

Longitudinal Trends:

Cohort Replicative Design

Year:	81	82	83	84
Reading	57	59	60	60
Math	58	61	62	64

Cohort Repetitive Design

Grade:	5	6	7	8
Year:	81	82	83	84
Reading	56	56	61	60
Math	51	60	64	64

CALIFORNIA:

Uses self-made test in Spring. Metric: Score on special scale. Grade 8.

Grade 8 testing began in 1984, results were not presented in reports available to us for review.

CONNECTICUT:

Testing program modeled after NAEP: not all content areas are tested annually. Reports on hand did not or math.

LOUISIANA:

No grades tested in range 7-9. Only two time points were covered in the 1984 report.

MAINE:

Self-made test (or NAEP) given in the Spring. Metric: Average percent correct. Grade: 8.

Sex contrast: 1982

	Male	Female	
Reading & Language Art	70.89	74.26	percent correct

The Technical Report sent to us did not present longitudinal data.

MINNESOTA:

Test given in Fall. (Source of items not clear). Metric: Average percent correct. Grade: 8.

Sex Contrast by year on TOTAL score:

	Male	Female
1977	74.5	78.8
1981	75.0	80.1

Longitudinal Trend:

	1977	1981
Comprehension of longer discourse	72.9	79.4

NORTH CAROLINA:

CAT, given in the Spring. Metric: Varies. Grade: 9.

Sex Contrast: 1984

	Male	Female	
Comprehension	56	63	National Percentile
Math Computation	56	67	

Longitudinal Trend:

Year	81	82	83	84	
Reading total	7.8	10.1	10.1	10.1	Grade equivalent

PENNSYLVANIA:

Self-made test given in Fall. Metric: Mean score. Grade 8. 1982 special report [school samples each year are volunteers, not a probability sample.]

Sex Contrast:

Reports available did not present sex contrasts.

Longitudinal Trend:

Year:	78	79	80	81
Reading	22.0	27.0	27.1	27.6
Math	32.0	31.8	32.0	32.5

RHODE ISLAND:

ITBS administered in the Spring. Metric: Median Percentile Rank. Grade: 8.

Sex Contrast:

Discussed in text, not tabulated. Direction of difference was mixed within and across grades.

Longitudinal Trend:

Year:	82	83	84
Reading Comprehension	51	60	56
Computation	52	55	60

SOUTH CAROLINA:
CTBS given in the Spring. Metric: Varies. Grade: 7

Sex Contrast: 1984

	Male	Female	
Total Reading	41.6	46.0	Percent above the national median score
Total Math	41.2	53.7	

Longitudinal Trend:

	1983	1984	
Total Reading	41.9	44.1	Median natn'l percentile
Total Math	44.5	51.7	

TEXAS:
Self-made test in the Spring. Metric: Percent mastering content. Grade: 9.

Sex Contrast: 1983

	Male	Female
Reading	77	83
Math	78	80

Longitudinal Trend:

	80	81	82	83
Measurement	70	69	76	79
Total Reading	70	69	72	80

VIRGINIA:
SRA Achievement Series in the Spring. Metric: ?
Grade: 8.

No sex contrasts were given in this report.

Longitudinal data were given for outcomes other than test scores.

WISCONSIN:
CTBS and self-made test given in Spring. Metric: varies.
Grade: 8. 1983 Report.

Sex contrasts were not reported in reading or math.

Longitudinal Trends:

	1980	1983	
Reading	71%	74%	Percent correct on self-made test

CTBS	76	77	78	79	80	81	82	83	
Reading	64	62	57	62	62	62	64	64	Natn'l
Math	72	70	59	61	66	66	72	70	%

Analysis of these data show that reports of state testing programs will not be a likely source of information on within-state contrasts that can be readily used to make state-to-state comparisons. Of the 13 states we examined more closely, four produced no trend or sex contrast and skill areas of interest. Among the remaining nine states, six presented sex contrast data and eight presented trend data.

Gender identification was one of the most frequently reported characteristics by state assessment systems (about one quarter of all states), yet we found only six states that reported sex contrasts in the most frequently tested skill areas and grade span. We concluded that subgroup data that are even roughly comparable across many states will be very hard to find in published reports. If the raw data could be obtained, it might be possible to produce subgroup contrasts in more states, but the coverage of the nation is likely to be sparce.

Longitudinal trend information was reported by substantially more state assessment systems (over half have data covering three years or more). However, when we constrained our exmaination to grades 7,8, and 9 in reading and math, only 60 percent of the reports gave longitudinal information. We estimate that only 15-20 state testing programs report trend data in reading and math in this grade span. In this case the archival data in the states could probably be used to create more within-state trends for comparative purposes, perhaps covering a significant fraction of the nation. The question would remain of how to interpret the results.

The trend information we found revealed generally stable to increasing scores. It is not possible to compare rates of increase, given the differing metrics of the results, however. We don't know how valuable this information would be to state or national policy makers. The national trend (in recent years) could be inferred to be stable or rising. But this does not reveal what sudents have actually attained, only that they are attaining as well as (or slightly better than) before. If trends in different states were very contrastive (negative vs. positive, since differences in rate cannot be judged on the basis of the reported data) over several years, it might lead to a search for explanatory factors.

The longest series, from Wisconsin, reveals the potential benefit of comparative data. If data from other states were also available for this span, it might be possible to tell whether the 1978 "dip" in Wisconsin was unique to that state or occurred in more of the nation. If it was unique, a further analysis of events in Wisconsin might reveal a plausible cause which could be subject of further study, and might serve as a warning to other states.

While within-state contrasts could contribute to a national profile of academic achievement as well as providing interesting comparisons among the states, the reports from state assessment systems do not, at present, contain enough information to make it possible to develop these contrasts in very many states. Longitudinal trends are reported more often than subgroup contrasts. The data bases on which the reports are based may contain additional information that could make more within-state contrasts of both types possible. If state officials could be persuaded that such contrasts would help them to interpret their assessment data, they might be encouraged to allocate more resources to reporting these analyses.

Summary and Recommendations

At the outset we had thought that it might be possible to develop Consumer Report-type within-state trend and subgroup contrast indicators from existing state data to provide an alternative basis for between-state performance comparisons. Our analysis indicates otherwise. The degree of conformity in practices across states is too limited to pursue the matter further at present.

We believe, however, that the types of auxiliary information collected in at least some states represent valuable sources of data that, if broadly collected, could provide useful contextual information in the interpretation of state comparisons. The idea of making between-state comparisons of within-state longitudinal trends and subgroup contrasts still has merit if the information were available. Moreover, the existing state testing program annual data collection effort is an efficient vehicle to gather auxiliary information to expand the set of context, resource, and process indicators.

If the decision is made to proceed with a States-coordinated effort to link existing state tests (e.g. through the CCSSO Assessment and Evaluation Coordinating Center), then we urge that the group responsible for coordinating the test linking activities also develop plans for obtaining a select set of auxiliary information on a routine basis. Thus, to encourage and facilitate the range and quality of information to be provided by states for comparative purposes, we recommend that

- cooperating states should be encouraged to provide to the Coordinating Center on an annual basis uniform documentation describing their data collection activities;
- cooperating states should work toward the establishment of a common set of auxiliary information about student and school characteristics to collect along with their testing data. A standard set of definitions for measuring the chosen characteristics should be determined; and
- as one of its activities, the Coordinating Center should consider ways of contextualizing the State test comparison data to mitigate against the possibility of unwarranted interpretations of comparative results. The auxiliary information gathered as part of the previous recommendation should contribute to this activity.

Chapter 6
Overall Summary and Recommendations

The results of the feasibility study conducted by CSE on using existing data collected by States to generate state-by-state comparisons of student performance have been described and discussed in this report. Specific chapters were devoted to descriptions and summaries of general characteristics of current state testing programs (Chapter 2), alternative approaches to linking test results across states to create a common scale for comparison purposes (Chapter 3), detailed content analyses of currently used state tests (Chapter 4), and the availability of auxiliary information about students and schools and its potential for use in generating within-state comparisons that could serve as between-state indicators of educational progress (Chapter 5). Each chapter was intended to focus directly on particular concerns that need to be resolved prior to a major effort to rely on state-developed data for comparison purposes.

The best answer to the question of whether state-level can be used for state-by-state comparisons "it depends." From the outset we knew, and through our examinations confirmed, that there is a substantial amount of pertinent information collected by the states. The characteristics of state testing programs are quite diverse. While there are concentrations of testing in certain grades during the spring, not all states operate testing programs. Furthermore, the specific components of state testing programs are not necessarily the same over time; in fact during the next few years, virtually every state will change its testing activities including some states who will conduct statewide testing for the first time.

For the most part, however, movements on the testing front are forward and expansionary, increasing the likelihood of overlap in testing conditions across states. Testing changes within states are driven by a variety of stakeholders but the same sets of stakeholders (legislators, governors, business groups, parents, universities) are participating virtually everywhere. If the tendency toward a common set of goals for state-level educational reform efforts continues, the conditions for cross-state comparisons of educational performance will improve. Right now we can say that such comparisons using state data are "potentially" feasible. Given likely future developments across the states and selected properly targeted studies of the effects of different testing conditions over the next several years, the operative adjective could shift to "probably"; by the end of the decade, the answer could be "definitely" or "not in the foreseeable future". It is simply too difficult to speculate about what might come to pass given current state activities.

Our response to the charge to the STQI Project has been to attempt to document current practice and to consider what could be done to improve the conditions for use of state data for achievement comparisons. Our recommendations focus on the conditions that would have to exist before the data from states could be compared, and on the steps that would need to be taken to implement cross-state comparisons. In the remainder of this chapter, we restate the recommendations derived from our investigations. The location of these recommendations within the earlier chapters is noted so that the reader can readily place them within the context of their justification and elaboration.

Preconditions and Guiding Principles

Several recommendations dealt with the basic conditions that should exist before using data from a state in performance comparisons and the principles that should guide the development of achievement indicators from state data sources.

ISSUE:
Which states should be included in cross-state comparisons?

RECOMMENDATION:

The comparison should include only those states where there is sufficient empirical evidence to allow analytical adjustments for the effects of differences in testing conditions. All states that collect test data on the pertinent content areas at the designated grade levels or whose test results can be statistically adjusted to the targeted testing conditions should be considered for inclusion in cross-state comparisons. (p. 3.2)

ISSUE:
What principles should guide the selection and development of achievement indicators derived from existing state test data?

RECOMMENDATION:

1. Existing state testing procedures should be disrupted as minimally as possible. Only those data collection activities considered essential for obtaining evidence of comparability should be introduced over and above the states' own planned expansions and extensions of their testing activities.

2. Existing state tests and testing data should be used as much as possible.

3. Regardless of the optimal specificity desired in the reporting of cross-state performance, the content of the tests to be used for comparisons purposes should be specified at as low a level (subskill or subdomain) as possible to enhance the quality of the match to existing tests and to encourage attention to the content and detail of what is being tested.

4. If the cross-state comparison are to be achieved through linking of a state's test to a common linking test, the content covered by the linking tests should be as broad as possible both to ensure overlap with each state's tests and to encourage

broadening rather than narrowing of the curriculum across the states.

5. The proposed approaches for developing state-by-state achievement indicators should be compatible with the wider issue of the development of systems for monitoring instructional practices as well as educational progress both within and across the states. Desireable augmentations of current state practices should increase documention of student and school characteristics within the framework of planned changes in state educational activities. (p.1.9)

Proposed Approach

At various times during the STQI Project, a number of approaches were considered for using equating and linking methodologies for placing different states' test results on a common scale for cross-state comparisons. The deliberations on these alternatives by project panelists and staff, along with input from other participants in panel meetings and other groups (e.g., CCSSO representatives), led to a recommended approach for linking state test results and recommendations for its implementation.

ISSUE:

What approach should be used to place state test results on a common scale?

RECOMMENDATION:

1. A common anchor item strategy, wherein a common set of linking test items is administered concurrently with the existing state test to an "equating-size" sample of schools and students, should be used as the basis for expressing test scores from different states on a common scale. (p. 3.7)

2. The items contributing to the common anchor set should be selected from multiple sources including existing state-developed tests, NAEP, commercially available tests, and other policy relevant and technically adequate sources, such as the IEA tests. (p.3.12)

ISSUE:

What additional issues should be considered in implementing the desired alternative for linking state tests?

RECOMMENDATIONS:

1. The mechanisms for establishing the skills to be included in the common anchor set, for selecting items to represent the skills, and for specifying the rules for participation by individual states should be developed and administered primarily by collective representation of the states. (p. 3.12)

2. The organization responsible for developing and administering the linking effort should consider the following points relevant to implementation:

a. Procedures for documenting contents of existing state tests should be specified so that questions of what is being equated to what can be addressed.

b. Specification of content represented in common anchor set should be at the lowest level possible (subskill level) even if achievement indicators, at least initially, are to be reported at higher levels (skill or content area).

c. The minimum criteria for considering an item for inclusion in the common anchor item set should include

- The item measures a skill selected for the common anchor item set, and
- Sufficient empirical evidence is available about the item to ascertain its behavior for the major segments of the student population with which it will be used.

d. The selection of items should be made by teams of curriculum and testing specialists from a broad-based pool of items without identification of their source.

e. The following set of testing conditions should be specified:

- Target grades and range of testing dates along with requirements for special studies in those states who normally test outside the chosen range or do not test at present but elected to participate.
- Procedures for concurrent administration of the common anchor item set with existing state tests for the various alternative types of state tests (matrix sampled,state-developed single form, commercially developed standardized test).
- Auxiliary information for checking subgroup bias and determining sample representativeness (for equating and scaling purposes).
- Minimum sample sizes (for both schools and students). (pp.3.13-3.14)

Pilot Study

Before proceeding with full-fledged implementation of any approach to achievement comparisons based on test data from existing state programs, project participants expressed the belief that the impact of deviation from targeted testing conditions should be studied further. The desire for empirical evidence about the consequences of the proposed alternative led to project activities designed to identify content areas and grade for an exploratory study of the proposed linking strategy.

ISSUE:

What additional information is desireable in order to determine whether it is practically feasible to link existing state tests?

RECOMMENDATIONS:

1. A pilot study of the proposed common test linking strategy should be conducted in a limited set of skill areas for a specific grade range in order to determine both the quality of the equating under preferred conditions and the effects of various deviations from these conditions.(p. 3.3)

2. The content areas and grade levels to be used in the pilot study should be literal comprehension for reading and either numbers and numeration or measurement for mathematics at grades 7-9. (p. 4.27)

Auxiliary Information and Documentation

Part of the project effort was devoted to determining what auxiliary information states collect and/or report about the characteristics of their students and schools and whether it might be possible to develop within-state trend and subgroup contrast indicators from existing state data to serve as an additional source of between-state performance comparisons. Our investigations indicated that while there is a wide variety of auxiliary information collected across the states, there is too little conformity in practices at present to make such comparisons viable. Nevertheless, the types of auxiliary information collected in at least some states represent valuable sources of data that, if broadly and uniformly collected, could provide useful contextual information for state comparisons. To encourage and facilitate the collection and reporting of common auxiliary information by the states, several additional recommendations were made.

ISSUE:

What steps should be taken to encourage and facilitate the collection and reporting of common auxiliary information about characteristics of students and schools?

RECOMMENDATIONS:

1. The organization responsible for coordinating the test linking activities described earlier should also develop plans for obtaining routinely a select set of common auxiliary information from states about their students and schools.

2. Cooperating states should be encouraged to provide on an annual basis uniform documentation describing their data collection activities.

3. Cooperating states should work toward the collection of a common set of auxiliary information about student and school characteristics along with their testing data. A standard set of definitions for measuring the chosen characteristics should be determined;

4. The organization responsible for coordinating test linking efforts should consider ways of contextualizing state test comparison data to mitigate against the possibility of unwarranted interpretations. The auxiliary information gathered as part of the previous recommendation should contribute to this activity. (pp. 5.17-5.18)

Political, Institutional, and Economic Environment

Most of our remaining recommendations regarding the implementation of the common test linking strategy had to do with the establishment of an effective political, institutional, and economic environment for the proposed indicator effort.

ISSUE:

What type of environment must be established if the proposed indicator effort is to be successful?

RECOMMENDATIONS:

1. To develop the necessary levels of political support for this activity, broad-based support for the idea should be developed. Key participants include Chief State School Officers, their staffs,and other state education officials; other prominent state officials, including the Governor, Members of Congress, and state legislators; and representation of members of large city school districts, the education associations and from the private sector.

2. An institutional structure for the conduct of this activity that relies heavily on the collective efforts of the states should be adopted. The Council of Chief State School Officers' new Assessment and Evaluation Coordinating Center proposal deserves consideration for this purpose.

3. Technical assistance and oversight should be established to assure the technical and methodological quality of the linking and equating, of the content of measures, and of validity of interpretations. This oversight should be provided by independent or semi-independent panels, perhaps modeled on the panels advising the NAEP activity.

4. A long-term, secure basis of financial support for coordinating and updating the test linking activity and the collection and reporting of common auxiliary information should be developed. This support is necessary to ensure that modifications in the basis of comparison and in the participating states can be accommodated over time while maintaining the integrity of the linking effort. (p.3.14)

Cost Implications: An Addendum

During the STQI Panel meetings and in subsequent discussions with federal and state personnel interested in education quality indicators, questions about costs of linking state data for achievement comparisons were raised. Although a cost analysis was not explicitly called for contractually, the possible cost implications of our proposed alternative is considered in a separate addendum to the report prepared by Darrell Bock (Appendix 20). This addendum lays out the basis for a small-scale feasibility study of the test linking option proposed and provides a cost estimate of approximately $80,000 (direct cost) assuming that approximately 3 schools from each of 5 states (with varying testing configurations) were to participate in the study.

Note that this cost estimate is for a limited pilot of one grade level in a few skill areas and assumes that states would bear certain of the routine field costs themselves. At the current stage, there is insufficient information to provide reasonable ball-park cost figures for a broader feasibility study at other grades with a wider range of skills or for full implementation of such a linking system. In our view there needs to be further discussion about possible directions of the state efforts in testing and on the desired level of effort toward comparable achievement indicators before such numbers can be reasonably generated.

REFERENCES

Baglin, R.F. (1981) Does "nationally" normed really mean nationally, Journal of Educational Measurement, 18(2), 97-108.

Harnisch, D.L. (1983) Item response patterns: Applications for educational practice, Journal of Educational Measurement, 20(2), 191-206.

Keesling, J.W. (1985) Identification of treatment conditions using standard record-keeping systems, In L. Burstein, H. E. Freeman & P. H. Rossi (Eds.), Collecting Evaluation Data: Problems and Solutions, Beverly Hills, CA.: Sage Publications, Inc., 207-219.

Neigher W.D. & Fishman D.B. (1985) From Science to technology: Reducing problems in mental health evaluation by paradigm shift, In L. Burstein, H.E. Freeman, & P. H. Rossi (Eds.), Collecting Evaluation Data: Problems and Solutions, Beverly Hills, CA: Sage Publications, Inc., 263-298.

U.S. Department of Education (1984) State Education Statistics: State Performances, Resource Inputs, and Population Characteristics 1972 and 1982.

APPENDIX 1

APPENDIX 1

PANELISTS FOR FEASIBILITY STUDY OF STATE TESTS AS QUALITY INDICATORS

R. Darrell Bock, Professor, Department of Behavioral Science and Education, University of Chicago

Dale Carlson, Director, California Assessment Program, State Department of Education

J. Ward Keesling, Advanced Technologies, Inc.

C. Thomas Kerins, Manager, Program Evaluation and Assessment Section, Illinois State Board of Education

Robert L. Linn, Professor, Department of Educational Psychology, University of Illinois, Champaign

Edward D. Roeber, Supervisor, Michigan Education Assessment Program

Richard Shavelson, Professor, Graduate School of Education, University of California, Los Angeles and Rand Corp.

Loretta A. Shepard, Professor, School of Education, University of Colorado

Marshall S. Smith, Director, Wisconson Center for Educational Research

APPENDIX 2

APPENDIX 2

Telephone Interview Guide
for
Quality Indicators Study

I. Introduction

1. Introduce yourself: Hello, I'm __________, from the Center for the Study of Evaluation at UCLA.

2. State Purpose of Call: We are contacting State Assessment Directors in regards to a study which we are conducting on behalf of the National Institute of Education (NIE) and the National Center for Educational Statistics (NCES). This study was prompted by a concern on the part of Chief State Officers about the development of appropriate indicators of educational quality at the state level. One of the sources of information which could possibly be used for this purpose is existing state assessment or competency data. The reason why we are contacting you, then, is to obtain some information about your testing or assessment program. We hope that based upon the information which we gather from all the state assessment directors that we will be able to provide recommendations about whether it is methodologically feasible and economically reasonable to use existing state assessment information as indicators of educational quality.

Before we begin, you should know that the study has the support and cooperation of the Chief State School Officers, as well as that of some of your colleagues such as Dale Carlson (California), Ed Roeber (Michigan), and Tom Kerrins (Illinois). We appreciate your cooperation and will provide you with summaries of what we eventually produce.

To facilitate these calls, we have organized our questions into three major sections: Overall design of program, reports, and data availability. In the initial section, overall design, we wish merely to confirm information which we already have and to complete any omissions. In the latter sections, some of the questions may be answered through documents which you could send us. If so, please indicate that and we will proceed more rapidly.

II. Overall Testing Program

Our records indicate that:

1. Does your state have a statewide testing or assessment program whose purpose is other than assessing the minimal competency level of students? Yes _____ No _____
2. Does your state have a statewide minimum competency testing? Yes _____ No _____

If the answers to both of the above were NO, then go to Question 6 at the end of the last section.

3. For each of the above, what areas are tested:

Assessment:	Reading	Math	Writing	Other________
Competency:	Reading	Math	Writing	Other________

4. At what grade levels are these tested:

Assessment:	Reading	Math	Writing	Other________
Competency:	Reading	Math	Writing	Other________

5. Are each of these levels tested annually, and if so what month(s)?
Yes _____ No _____

If No, on what basis are they tested?____________________________
____

6. Now we would like to understand your student sampling strategy:

Do you test all students at a grade? Yes____ No____

If No, please describe your sampling:____________________________
____

7. For what purposes are these tests used:____________________________
____

8. Are the test items developed internally ____ or externally ____.
If externally, who developed them ______________________________?
Name of test ____

11. Are there embedded external items? If Yes from where ____________

9. Are you aware of other states that use the same or some subset of the same items?
Yes ____ (Specify which: ______________________) No ____

10. Are you planning any major changes in the program for next year?
Yes ____ No ____.

11. What about changes planned beyond next year?

BEST COPY AVAILABLE

III. Reports

Now, we would like to switch our focus to the reports which your program regularly prepares and which are generally available.

1. Do you produce the following types of reports for your program:

   ____ Technical Reports, describing Psychometric Properties of the tests.

   ____ Content Reports, providing Content Specifications.

   ____ Analysis Reports, providing summaries of the results.

2. Can we obtain copies of these reports. Yes ____ No ____

3. What is the most recent school year for which these reports are available? Year ____.

4. In your Content Reports, do you provide the following:

   ____ Objective Statements

   ____ Domain Specifications

   ____ Sample Items

   ____ Description of Test Construction Procedures

   ____ Description of Item Sampling

5. In the Techincal Reports, do you provide information about the following:

   ____ Sub-Group Differences (Specify types of information reported)
   ____

   ____ Item Characteristics (Specify types of information reported)
   ____

   ____ Reliability (Specify types reported)
   ____

   ____ Content Validity (Specify types of information reported)
   ____

   ____ Construct Validity (Specify types of information reported)
   ____

   ____ Predictive Validity (Specify types of information reported)
   ____

BEST COPY AVAILABLE

6. We are particularly interested in all your reports which contain results from the tests. The following questions all concern these reports.

a. Could your briefly ennumerate the reports that contain results that you regularly produce (other than reports back to the schools and districts, though we would like to receive sample copies of these):

____

____

____

b. In these reports, are the results provided for a single year? _____
Or, do you provide longitudinal or time trend data? __________
If the latter, for what periods? __________

c. What unit of analysis do you use in these reports: school, district, state? __________

d. Are the results reported in the aggregate for the whole state? _____
Or, do you report results for subgroups, e.g., by sex, race, socio-economic language, community type. _____

e. If you report results for subgroups, what characteristics do you use to define those groups? __________

____

____

f. When you report the results, what type of scale do you use?
____ percentiles
____ number correct
____ scale score
____ percent correct
____ other (Specify): __________

g. When you report the results, generally what form of statistical summary is provided:

Measures of Central Tendency (Specify which)

____

Measures of Dispersion (Specify which)

____

Frequency Distributions (In what form:)

____

Proficiency Levels (percentages passing or reaching criteria)

____

Other, Please describe:

____

____

____

BEST COPY AVAILABLE

h. Are these statistics provided for all subgroups? ____________

i. What statistics or method of presentation do you use for longitudinal data? ______________________________

____

7. Are there other reports which you produce that contain results or information about the educational quality in your state? ____________

____

____

____

IV. Data Availability

One of the avenues we are examining is whether it might be feasible to actually use and reanalyze state assessment data in order to derive indicators of quality. Therefore, we would like to know about the data which you collect from the tests.

Would the data you have collected from your test be available for analysis by us? Yes ____ No ____ (go to 6) Maybe (Specify the conditions: ____).

If yes, what are the procedures for obtaining the data? ______

____

____

How long will it take? ____________________

How much will it cost? ____________________

2. Is the data available on computer tapes? Yes____ No____, Specify ________

3. Is the data stored at the student level? Yes____ No____

4. Is data available at the item level? subtest? total test?

____

5. Besides test scores, what additional information is stored at this level (i.e. race, sex, etc.)? ____________________

____

6. Other than testing programs at your state, is other information collected by the state which might be used for this study? (indicators of quality or indicators of context)

   Yes ____ No ____ (If No, go to end.)

   - What agencies house this information: ____________________

     ____

   - Could you please identify appropriate contact people at these agencies: ____________________

     ____

   - What type of information is available? ____________________

     ____

     ____

     ____

   - What is the approximate volume (number of students, etc.) of the existing data? ____________________

BEST COPY AVAILABLE

- Is it available in reports? (If so, please indicate titles):

____
____
____
____
____

- Is it available in computer compatible format? Yes____ No____

END: Thank you for you help with this project. As I mentioned at the start, we will provide you with a summary of results at the end of the project.

Addendum

USE PURPOSE OF STUDY

This proposal responds to the request for proposal issued by the National Institute of Education for a feasibility study of the use of state tests as indicators of educational quality on a national level. The study will address whether existing state tests may be combined to give a picture of educational effectiveness. The ultimate goal of the research will be to provide a better database for judging educational policy. The technical approach of the stusy will draw upon statistical, psychometric, and policy expertise to determine the feasibility of reaching this goal.

APPROACH

Using a panel of technical and policy advisors as well as consultants with special expertise, CSE will conduct a feasability study of the various alternatives jointly suggested by these indicators. Thus, the initial task of the study is to identify the range of alternative approaches and their respective technical resource requirements. In addition, CSE will conduct a survey to determine the nature and extent of existing assessment data in each state. Using the results of this survey and an examination of the materials obtained from the states, CSE will prepare a set of recommendations regarding the relative technical economic feasibility of the different alternatives. These results will be received by theAdvisory Panel and their recommendations and suggestions will form the basis for the formal project report.

SCHEDULE

The project was initiated at the beginning of October, with two Advisory Panel meetings scheduled for late November and January. The formal report will be available after the last panel meeting.

APPENDIX 3

APPENDIX 3

Revised 5/15/85

"State Tests as Quality Indicators" Project
Center for the Study of Evaluation
Partial Summary of First Policy and Technical Panel Meeting
Washington, D.C.
November 29-30, 1984

The first Policy and Technical Panel meeting for the State Tests as Quality Indicators (STQI) Project, being conducted by the Center for the Study of Evaluation (CSE), was held at the National Institute of Education on November 29-30, 1984. While attendance at the meeting fluctuated, the participants included representatives from the following organizations and agencies: National Center for Educational Statistics, National Institute of Education, CSE project staff, STQI Project Policy and Technical Panel members, National Assessment of Educational Progress, Office of Planning, Budgeting, and Evaluation of the Department of Education, and the National Association of School Boards of Education.

CSE Statement of Objectives of the Project and the Panel Meeting

The 11/29 meeting began with a discussion of the overall objectives of the STQI Project and of the first panel meeting. The overall project objective is to explore the feasibility of using equated or aggregated state testing results as national or state-by-state indicators of educational quality. This exploration is to entail a documentation of existing state testing program activities with specific emphasis on the possibility of using data already routinely collected to form "comparable" state-level indicators and, if so, to determine the types of analytical and psychometric methods necessary or potentially appropriate to generate the desired indicators. With respect to the latter, the original CSE proposal had identified essentially four general approaches to derive indicators using state data: equating of test content; econometric adjustment for selection and/or economic and socioeconomic conditions; equating by the use of a common test or linking measure; and methods that depend only on within-state information such as trend data and subgroup comparisons.

The purpose of the first panel meeting was to consider which of the available approaches for deriving indicators from state data were potentially useful given current testing practices, and thus which approaches CSE should explore in greater depth using reports provided by the states. As part of the preparation for the meeting, CSE conducted in-depth telephone interviews with representatives from state testing programs and requested copies of existing reports and content specifications generated by the state testing programs. The results of these phone interviews were then combined with information from other recent surveys of state testing activities and distributed to meeting participants. It was expected that this information would place the proposed approaches within a context of existing practices and aid in the effort to refine and focus the remaining tasks of the feasibility study.

Federal Perspective on the Project

A brief statement of the federal perspective on the intent of the project was then provided by Emerson Elliott. In his remarks, Elliott placed the present project within the context of recent federal initiatives on educational indicators. These initiatives are most directly reflected in Secretary Bell's 1984 release of the State Education Statistics Chart and the work of the Department of Education's Indicators Project. Their intent, along with the support for the STQI project, is to provide national and state-by-state data that help to answer three questions. Namely,

1. What is the health of American Education?
2. What are students learning?
3. Are things getting better or worse?

Director Elliott indicated that he did not believe that the attempt to address the above questions using state-level data as quality indicators necessarily meant that the states must be ranked. Within-region comparisons and longitudinal patterns within states were cited as examples of other types of information that would serve to inform policy makers with respect to the major questions of interest. What is of primary interest is the compilation of a national picture of what's happening in the states with respect to the quality of their educational programs.

Elliott's specific expectations for the STQI project had shifted somewhat from his original objectives. Early on, he had thought that this project might yield some indicator data that could conceivably be included in the next (1985) release of Secretary Bell's chart. However, given the accelerated time-line of the new chart (to be released in December 1984), this goal no longer was reasonable. Moreover, given a new awareness about the diversity of the existing state testing programs and the broad-based changes in these programs that have recently occurred or are currently in progress, it does not appear likely that existing state testing activities can readily serve as a means of generating comparable and stable indicators of educational achievement across the states in the near term. And, given recent actions by the NAEP Policy Panel and council of Chief State School Officers (CSSO), it may be possible to generate state-level NAEP performance indicators in about five years. If this were to occur, there might be less long-term interest in using state testing data as indicators.

Given the changing situation, Elliott ultimately would like the STQI project to provide further insights into whether the assessments states administer and report for their own use can be synthesized to form indicators of national trends in educational quality. In addition, he hoped that the project could contribute material for a section on national achievement to appear in the revised Indicators Reports to be published periodically by the National Center for Education Statistics (NCES).

At this point, participants cited other activities on education indicators that were related to either the federal initiatives or this project's efforts. Other agency and organization work mentioned included the National Academy of Science Project on Mathematics and Science Indicators funded by the National Science Foundation, relevant sections from the General Accounting Office's examination of the National Science Board's Report on the Status of Science Education, (CCSSO's recent vote in support of developing state-level education achievement indicators that might be used for state comparisons and their efforts to build state and national capacity for collecting data on other areas relevant to education achievement, and possible activities by Congress and the Office of Technology Assessment. There was a general sense of movement across a broad front to develop a national capacity to collect and report information that may serve as indicators of the quality of the American education system.

Description and Discussion of Available State-Level Information

The available results of CSE contacts with representatives from state testing programs and examinations of reports from other sources regarding these programs were described. Copies of the Telephone Interview Guide for the first round of calls to state testing programs (Attachment I), a draft version of a chart containing state-by-state responses to key sections of the interviews (Attachment II; note that this chart has been updated since the meeting to reflect additional state contacts) and a brief summary of selected facts about state testing programs (Attachment III) were distributed and discussed. The general consensus of participants was best reflected in the comments from the state testing program members of the panel. They agreed that the handouts clearly reflected accurate information about variations in existing state programs, but that the actual picture was even more complex than depicted. Our interview data apparently represented testing programs fairly, particularly the kinds of tests and the testing targets (minimal competence, basic skills, broadly measured achievements, exceptional educational performance). Less well detailed was the function these tests were designed to serve and how they are currently used. All programs are subject to change but that change has accelerated, largely as a result of state reform initiatives in response to the National Commission on Excellence in Education report.

The discusssion at this point also touched on a number of other issues and ideas briefly, including the possibility of subgroups and/or content disaggregation of state test results, the variation in the timing of testing programs, the desirability of a quality indicator for state-level longitudinal and subgroups trend data patterned after the Consumer Reports automobile indexes or the Consumer Price Index, questions regarding the commonality of content across states, the potential for use of shared item banks, and better coordination and cooperation with commercial test publishers.

Equating of Test Content

The discussion then shifted to direct consideration of the different methodological approaches for aggregating, equating, or otherwise combining measures as identified in the CSE proposal. The first approach considered was the equating of test content. This approach focuses on the content of state tests -- content specifications, items, subtests -- and considers whether it is possible to classify items on some basis (e.g., commonality of domain, difficulty) into "equivalent clusters" and then compare across states based on performance on equivalent items. The general trend of the discussion regarding this approach was that while it might be theoretically possible to equate on content, in practice a considerable number of complications exist making the notion impractical at present. Among the points made by participants were the following:

1. Not all states operating internally developed programs are equally conscientious about developing content specifications for the generation of test items.

2. Even among the states that do provide detailed content specifications, the match of test items to specifications and the distribution of items to objectives may be uneven.

3. To do a proper analysis of the content of state assessments as a first step in the equating, one could not simply rely on the content specifications. It would be necessary to examine actual items and tests and perhaps talk to the people who put the test together. The actual process of generating items tends to be an iterative interplay among the specification, the examination of the wording of each item, and the item statistics.

4. The level of abstraction that can be used to equate content is of concern. It may be that content equating is only feasible at the most general level (e.g., reading, math).

5. If one attempted to equate at too high a level of content specificity, the number and nature of items that qualify as common topics across states can artificially truncate differences in achievement.

6. It may also be important to remember that, in practice, in order to be able to combine items to form a score for comparison, one needs similar items given in essentially the same format (e.g., not vertical vs. horizontal) at roughly the same administrative time to the same grade under the same set of external sanctions with respect to performance (i.e. consequences of the performances). It may simply be impossible to satisfy all these conditions with existing state testing programs.

7. If their current interest in state-level NAEP data continues or expands, then the question of the match of the content emphasis of the state testing programs with that of NAEP is worthy of further consideration. (The same can be said for comparison with commercial tests in states where a specific publisher has a substantial portion of the local testing market.)

   Such efforts might provide a basis for the development of a national indicator with respect to the diversity of content of testing programs across the states.

With respect to the possible further work of the STQI Project on approaches emphasizing equating of test content, the discussion was summarized as follows:

1. There were substantial doubts about the utility of equating of test content.

2. There was some support for providing a more in-depth description of the content of the state testing programs.

3. The possibility of developing an indicator of the diversity of the content of state tests warranted further examination.

4. If, even given the above, one still wanted to equate content, it would be necessary to work at some level of general objectives (perhaps more specifically than reading comprehension).

Participants raised several additional points that emphasized a need to go beyond an examination of test content to generate indicators of the school curriculum. There was interest in more direct quality indicators of curriculum activities at the state level. This interest suggests the need for thought about how to characterize the core objectives students are supposed to know and how to go about ascertaining this information. Several panelists cautioned about inferring what a state teaches based on what it tests. There was no indication that participants expected the STQI project directly to address these concerns; however, it was clearly perceived that any attempt to use content comparability as an indicator must be balanced against the potential for limiting the representativeness and validity of any such indicator as a measure of state-based activities and goals.

Econometric Approaches

The term "econometric approaches" was used to characterize procedures which involved attempts at analytical adjustment of state testing data to bring about a greater degree of comparability across states with respect to economic and socioeconomic factors as well as to the nature of the students within the state who take a given test. These approaches fall into two broad categories. In the simpler category, state test data are directly adjusted or weighted for a set of economic factors (e.g., state unemployment rate and other indicators of the health of the state's economy) or socioeconomic factors (e.g., indices of poverty, ethnic make-up, bilingualism) to arrive at a set of measures that presumably Compensate for these sources of non-schooling influences on educational achievement prior to any effort at cross-state comparison. The overall intent of such a strategy would be explicitly to take context factors into account in reporting state education outcome indicators.

The second category of econometric approaches would entail employing modern methods for adjusting for sample selection bias. Presently any attempts at using SAT, ACT, ASVAB or other non-census testing that occurs in multiple states as indicators is limited by the non-random and non-comparable sample of students within a state who take these tests. If it were possible to obtain student-level data on these tests and on the "pertinent" characteristics of the students who take the tests, in theory, it may be possible a) to apply selectivity modeling methods to adjust test performance for non-random selection at the student level (within and across states) and b) then to use the state-level aggregated adjusted scores as a basis for equating or linking the state testing program data. This strategy entails several strong assumptions about available data and, even under the best circumstances, may yield results with only limited precision.

Overall, participants were skeptical about the practicality of these types of adjustment strategies at the present time. With respect to the first category, there were questions about whether most states collected the right data in comparable ways in a sufficiently accurate manner.

Moreover, there were doubts about whether it was reasonable to expect to arrive at a mutually agreed upon (both technically and politically) set of weights or adjustments to achieve comparability of context or even whether this was a desirable goal. In addition, these kinds of adjustments do not remove the necessity of having to find some means of representing state test data on a common scale. That problem would still have to be solved. And, with respect to the possibility of using selectivity modeling to arrive at a common scale, the participants did not feel confident at this point about such a strategy if only because too little is known about how these methods work in practice.

Common Scale Through Equating and Linking

Robert Linn introduced the discussion of using equating or linking of tests by providing a historical perspective on other efforts pertinent to this task. He described the Anchor Test Study which attempted to equate commercially published standardized achievement tests for the purposes of Title I evaluation. It was pointed out that this study required substantial resources and time and its value quickly deteriorated as publishers modified their test content and renormed their tests. Linn also discussed the problems with the TIERS (Title I Evaluation and Reporting System) data that still remained even after the Anchor Test Study and subsequent development of NCE scales and TIERS evaluation models. In addition to remaining equating errors, the strong effects of time of year for testing and test administration conditions were cited.

Finally, Linn briefly discussed the question of NAEP as a common scale for state comparisons. While this is an obvious possibility that will attract further consideration, he reminded the participants that NAEP tests contain only small numbers of items on any objective and may not represent all content of interest for inclusion in state outcome indicators.

Darrell Bock then presented the basic psychometric alternatives for equating and linking state tests. Two strategies were described. The first strategy involves the use of common anchor items. This strategy requires that a set of anchor items (taken from NAEP, or from a pool of items provided by different states) be included on all tests to be equated and that item response theory (IRT) methods be used to scale these items within each state's tests. This strategy assumes the absence of any type of context and location effects for item placement within a test, of effects of time of testing within a school year, and of test administration conditions. Many participants were skeptical about whether such assumptions were practically justifiable.

The second strategy requires using matched data from students who take both the state-administrated test and the test chosen to serve as the anchor. This strategy could potentially be employed in states which have every pupil take the state test since students who took, for example, the NAEP that year would be doubly tested. To employ this strategy, it would have to be possible to match students person by person (i.e., students' NAEP scores with their assessment scores). To make this practical for statecomparison, NAEP would have to test more densely in most states (Bock estimated that it would require approximately 1000 matched kids at a given grade level to have any confidence in the IRT equating and scaling). There would also have to be enough information to adjust for time of administration of tests.

According to Bock, the possible advantages of this matched test data strategy is that random, representative samples of a state's student population are not required and a smaller sample of NAEP testing than would be needed to use NAEP itself as a state-level achievement indicator might suffice to use the NAEP as a benchmark. Also, the same strategy could be used in states where a commercially published test is given to a sufficiently diverse set of schools and students. Finally, such tests as the SAT, ACT, or ASVAB could be used to check the calibration.

The general discussion with respect to the matched test data strategy first focused on its costs and accuracy relative to alternatives such as state-level NAEP testing. Archie LaPointe described the current expanded state testing using NAEP (200 students each taking the same booklet, being conducted in Florida, Georgia, and Tennessee) and what size of study in a state would be required to fully implement a smaller NAEP (a so-called Mini-BIB with 7 booklets requiring 5000 students per grade level; roughly 100 schools in each state per grade level to yield at least 1000 kids taking 25 test items). There was some concern for BIB spiraling effects that the new NAEP procedure would introduce. Under Bock's proposed scheme, there presumably would be lower costs and fewer analytical complications than for the Mini-BIB design envisioned by NAEP.

The discussion then turned to the possible complications in employing a matched test data strategy and what kinds of information would be needed to decide whether to implement fully the approach. It was pointed out that the approach assumes one can obtain accurate estimates of individual abilities. Also, it would be necessary to calibrate the test items repeatedly because of possible item parameter drift. The concerns about test administration conditions and time of testing would still exist. One panelist cited the impossible tangles such a strategy poses, especially since it was to be done retrospectively.

The question was then raised about whether the utility of the matched test data strategy would depend on whether one wanted to compare a state's local objectives and performance nationally or to compare states on national objectives and performance standards. Two possible state-level advantages for linking to NAEP were identified, namely, 1) the local and state pressure to compare states to national norms and to other states, and 2) maintaining a certain degree of state control over tests. In the final analysis, it was agreed that the problem required states to grapple with the issues of the face validity for various stakeholders (state testing directors, CCSSO, legislators, Governors, public) of three alternatives: no common scale, equated scale, common test.

In order to decide which alternative is best, we need more information on the following issues:

1. Will a NAEP state-by-state mini-assessment yield more than just a total reading and math score? Would we also be able to provide urban/rural, regional (within state) and SES comparisons?

2. Would it be possible to pilot the matched test data strategy using existing data? Seven states (California, Florida, Illinois, Massachusetts, Michigan, New York, and Texas) currently have approximately 1000 students taking NAEP. What are the time and cost estimates of piloting this strategy in a cluster of states without additional data collection?

3. What additional new data collection would be necessary to have sufficient data to implement the matched test data strategy in a substantial number of states? What are the time and cost estimates for the expanded, full implementation version of this approach?

There was some discussion about whether there were other linking vehicles besides NAEP, in particular the possibility of doing such equating with commercially published tests. Several participants cited the possibly shifting attitudes of commercial publishers toward greater cooperation with NAEP as evidence of potential connections, and the substantial testing already being carried out in some states (e.g., CTBS and CAT used as state tests and by a large number of districts in some states which don't require it) which makes the use of commercial tests as a common link at least technically feasible.

In general, there was a consensus that the STQI project should devote further effort to identifying and describing the conditions states would have to meet to develop a common scale by using an anchoring approach of either type described above. This examination would presumably focus on technical consideration (timgin, dimensionality characteristics of the test, sample size needed) and resource and time considerations.

Within State Trends

The last approach discussed involved attempts to rely strictly on within-state data to yield cross-state comparisons. Operationally, this approach might entail developing indicators of longitudinal trends in performance within the state or subgroup (e.g., rural/urban, SES, ethnic or other student and school contextual characteristics) comparisons (either cross-sectionally or over time). If there were a sufficient number of states collecting: a) comparable data over time and b) comparable information that would allow disaggregation of test performance to the level of identifiable subgroups of students and schools, performance indicators based essentially on effect-size estimates (e.g., the year-to-year gains or urban-rural differences expressed in standard deviation units) could potentially be developed.

There are several potential problems with the within-state trends approach. The within-state comparisons would provide indicators of trends but not levels (relative versus absolute performance). Any changes in tests over time would potentially affect the validity of the longitudinal comparisons. Also, while states might nominally collect the same information relevant to classification by important subgroups, operationally,the specific measure of a given characteristic (e.g., definition of an urban versus a rural school, measurement of SES, ethnic and language classifications) used by states may differ sufficiently to hinder seriously attempts at cross-state comparison on this basis. The analytical model that would underride such indicators (i.e., choice of standard deviation to serve as the base for the effect-size estimate, and the model of normal growth underlying longitudinal trend measures) would also require further thought.

The consensus recommendation of the meeting participants with respect to the within-state trend approach to educational indicators is that this approach warranted further examination in the hopes that it may be feasible to derive a Consumer Report- type up-down trend indicator to include along with other achievement indicators that more directly reflect absolute levels of performance.

Closing Discussions and Suggestions for Further Work

In the final discussions, Emerson Elliott expressed interest in obtaining information about where state testing is going over the next five to eight years. He also hoped that the project might be able to provide guidance about the value and feasibility of developing indicators based on longitudinal series (with regard to test performance), the diversity of test content across states, and other indicators of the uniformity of state performance and educational program characteristics. There was general consensus that CSE should proceed with at least the following tasks:

1. Complete the interviewing about state testing activities and develop a chart that characterizes these activities.
2. Continue to obtain representative reports generated by state testing programs and conduct an analysis of their content with respect to the methodology used to develop, analyze, and report data at the state level.
3. Conduct an examination of the content of state tests including analysis of both content specifications and actual items where feasible.
4. Explore further the feasibility of developing summary Consumer Report-type indicators of trends with respect to diversity of content measures, complexity of skills measured, longitudinal changes, and subgroup differences.
5. Attempt to provide resource and time estimates necessary to both pilot and fully implement the approaches judged to be fruitful to arrive at state-level education indicators.

APPENDIX 4

Decision Memorandum on the Feasibility of Using State Level Data for National Educational Quality Indicators

Eva L. Baker and Leigh Burstein, Center for the Study of Evaluation,UCLA

Background

The desire for a national picture of educational quality remains a continuing but unresolved goal. Past efforts using available data from college admission tests have provided one source of information, but have been criticized because they represent performance of only one segment of the student population. Results from administrations of achievement measures of the National Assessment of Educational Progress (NAEP) provide a partial picture, but are limited because of the general character of the measures and the schedule upon which they are administered. Furthermore, because of NAEP sampling practices, no state by state comparative data are possible.

In the past, there has been some resistance from States about comparative information of any sort. The arguments have centered on the need for good contextualization of information so that differences in performance can be properly attributable to quality of educational services and not to social and economic conditions in the regions themselves.

A national test has been proposed periodically as a solution, but has been rejected because of the constitutional delegation of educational responsibilities to the States and the attendant notion that such a test would exert untoward Federal pressures toward uniformity in educational practices. The cost of such a new test (or radical expansion of the NAEP sampling and scheduling) would also be high.

Last fall, a question was raised among high level policymakers regarding the feasibility of using existing mechanisms within the States to contribute to the picture of American educational quality. Specifically under consideration was the extent to which existing measures of student performance collected by the States could be combined to 1) provide a national profile of performance in achievement domains; 2) provide a basis for state-by-state comparisons of student performance. A feasibility study was contracted to the UCLA Center for the Study of Evaluation (CSE) to explore the methodological and implementation issues of such an approach. This memorandum represents a summary of these analyses and recommendations regarding the feasibility of this approach.

Feasibility Study

A panel of scholars and practitioners was convened to engage in discussion of these issues. A list of participants is appended. These meetings were held in Washington, D.C., and were open to interested observers from government and professional organizations. Following the first meeting, CSE staff and the panel members developed options, collected information, and distributed preliminary findings. At a second meeting this spring, a general consensus was reached.

Methodological Issues.

The group considered a range of methodological options for combining State-level data for national comparative purposes. Opinions converged on using a common test linking and equating approach based on the administration of relevant common measures along with each state's own test to a sample of students.

Two concerns needed to be addressed before a decision could be reached about how this linking strategy might be applied. First, the question of possible content of the common tests was raised. To that end, CSE staff prepared a content analysis of tests or specifications of tests from 38 responding states who were conducting testing programs as of Spring 1984. The results of this analysis are included in our larger report. Based on our findings, the panelists recommended that two or three skill areas at a single grade level be chosen for initial examinations of equating options based upon the frequency of the skill areas' inclusion in State measures and the frequency at which various grade levels were represented in State test administrations. The areas of literal comprehension in the reading achievement area and either numbers and numeration or measurement in the mathematics achievement area at grades 7 through 9 were considered most suitable for initial equating efforts.

The second concern was the nature of the common measure proposed to serve as the basis for equating the disparate state measures. It was determined that technical procedures now exist that make it possible to equate tests without requiring that all sampled students respond to the same set of common items. However, the measures needed to share certain technical characteristics with the target measures in reading and math. Principal among these characteristics was unidimensionality of the scale.

Options

Various options were considered for the common linking measure. These will be briefly described below with a statement of their benefits and limitations.

Option One: Using NAEP measures for equating purposes

Benefits:

1. Measures exist
2. Measures have been developed with appropriate technical expertise

Limitations:

1. NAEP is not administered on an annual basis. Most State measures are administered annually and the goal of the Quality Indicators effort is annual reporting. Therefore, NAEP schedules might be changed at a significant cost, or the equating would become intolerably imprecise if "old" NAEP measures were used in between NAEP administration periods.
2. The current density of NAEP sampling does not provide a basis for equating in most states. NAEP sampling could be augmented, which would increase administration costs and would entail certain difficulties in interpretation of longitudinal data.
3. NAEP and state tests would have to be available from the same sample of students, at the same point in time. If NAEP schedules were adjusted to concurr with state testing schedules, then the NAEP data might not blend with the established NAEP testing schedules. If the state testing dates were altered to correspond to the NAEP dates, then data from the sample schools might not be equivalent to data obtained as part of the regular state testing effort.

Option Two: Creating a common pool of items drawn from existing State measures for use in equating

Benefits:

1. Measures exist (either State developed or publisher provided) and have empirical data associated with them.
2. Because measures would be derived from tests already used by States, they would more adequately reflect at least some local goals .
3. Cooperation and contribution to the pool would encourage State capacity building and the exchange of technology from States with better developed testing programs to those in relatively early stages.

4. Skill and content areas for equating would not be limited to current NAEP content areas, but could be developed based upon the actual interest and distribution of tested topics.

5. Costs for data collection would be low because the measure would be integrated with normal State testing practices.

Limitations:

1. This approach is dependent upon State cooperation. This cooperation in turn depends upon the political climate and local pressures upon a Chief State School Officer and the State testing program's operations.

2. Pilot studies would need to be conducted of the test pool used for equating on any skill or content area.

3. An organizational structure would need to be created to oversee this process and to assure technical and political sensitivity of the approach.

4. Assuming a successful trial period, some regular source of financial support external to individual States will be required.

Recommendation:

We recommend that the State item pool strategy be tried on an exploratory basis for a two year period, after which judgments about continuation, modification, or expansion could be made.

Implementation Issues Relevant to the Recommendation

It will be a serious matter to develop the necessary levels of political support for this activity. Key participants are, of course, the Chief State School Officers, their staffs,and other State education officials, but other prominent State officials, including the Governor, Members of Congress, and State legislators may need to be involved. Representation of members of large city school districts should be participants as appropriate. Broad based support for the idea should be developed.

Secondly, the matter of developing an institutional structure for the conduct of this activity should be considered. The benefit of having an organization of States manage the process will avoid the specter of Federal directive, and the Council of Chief State School Officers' Assessment and Evaluation

Coordinating Center proposal deserves consideration for this purpose.

Third, it is essential that technical assistance and oversight be established to assure the quality of technical and methodological operation of the equating, of the content of measures, and of validity of interpretations. This oversight should be provided by a panel, perhaps modeled on the panels advising the NAEP activity.

Fourth, a long-term, secure basis of financial support for this activity should be assured. The costs will not be high but resources should be regularly available.

Additional Technical Comments

Our interviews with State testing officials and examinations of reports and tests currently provided by individual states indicate an extensive range of activities of varying sophistication and quality. Many states collect and/or report a wide of array of auxiliary information about their students and schools along with their test data. Some states maintain and report longitudinal trends, and a few provide within-state comparisons, cross-sectionally or over time, broken out by major student and school sub-groups (e.g.,student sex, school size, type of community). These auxilliary indicators also represent valuable sources of data that could provide useful contextual information in the interpretation of state comparisons. The group coordinating the State Item Pool could be responsible for developing strategies for obtaining this ancilliary information on a routine basis.

To encourage and facilitate the range and quality of information to be provided by states for comparative purposes, we make the following additional recommendations.

- Participating states should be encouraged to provide on an annual basis uniform documentation describing their data collection activities (along the lines currently provided through the Education Commission of the States and the Roeber Survey).
- Uniform standards for documentation of the contents of State-administrated tests should be established. In the case of states using existing publisher-provided, standardized tests, the publishers should be responsible for providing the report to the state for transmittal to the coordinating center.
- Cooperating states should work toward the establishment of a common set of auxiliary information about student and school characteristics to collect along with testing

data. A standard set of definitions for measuring the chosen characteristics should be determined.

- As one of its activities, the coordinating center should consider ways of contextualizing the State test comparison data to mitigate against the possibility of unwarranted interpretations of comparative results.

A critical caveat is that these recommendations relate to State testing systems that are changing significantly. We believe that these changes, toward testing more students, more grade levels and more subject matters, will facilitate the capacity of State testing systems to contribute to a fuller national picture of educational quality.

APPENDIX 5

SOURCES OF INFORMATION ABOUT STATE TESTING PROGRAMS

1. Center for the Study of Evaluation, "Results from the Survey of State Testing Programs for the Quality Indicators Study," based on telephone interviews conducted November 12-26, 1984.

2. Southern Regional Education Board, Measuring Educational Progress in the South: Student Achievement, Atlanta, GA, 1984.

3. Roeber, E.D., "Large-Scale Assessment Programs: Program Descriptions, Summer 1984," Lansing, MI: Michigan Department of Education.

4. Roeber, E.D., "Survey of Large-Scale Assessment Programs: Fall 1983," Lansing, MI: Michigan Department of Education.

5. Anderson, B., "Status of State Assessments and MCT," Denver, CO: Education Commission of the States, March 14, 1984.

6. Pipho, C., "State Activity: Minimum Competency Testing," (Contained in Anderson, March 1984), Denver, CO: Education Commission of the States, January 1984.

7. Council of Chief State School Officers, "A Review and Profile of State Assessment and Minimum Competency Programs, 1984."

8. Pipho, C. and Hadley, C., "State Activity Minimum Competency testing as of December 1984," Denver, CO: Education Commission of the States.

9. Anderson, B., "Current Status of State Assessment Programs as of December 1984," Denver, CO: Education Commission of the States.

APPENDIX 6

QUALITY INDICATORS SURVEY SUMMARY

APPENDIX 6

General Characteristics

STATE	TESTING Have Program Y N	Used For: State Assessment Y N	Competency/ Proficiency Y N	No. of Testing Programs	ASSESSMENT PROGRAM: Areas Tested: Reading, Math, Other	Grade Levels	Selection Census Sample C S	Name of Test	Source of Items Internal External I E	Shared Items Y N	Major Planned Changes Y N
AL	Y	Y	Y	2	R,M,O	1,2,4,5, 7,8,10	C	SAT	E	N	Y
AK	Y	Y	N	1	R,M	4,8*	C		I	N	Y
AZ	Y	Y	Y(L)**	2	R,M,O	1-12	C	CAT***	E	N	Y ****
AR	Y	Y	Y	2	R,M,O	4,7,10	C	SRA	E	N	Y
CA	Y	Y	Y(L)	2	R,M,W,O	3,6,8, 10,12	C		I		Y
CO	N	-	-	-	-	-	-	-	-	-	-
CT	Y	Y	Y	3	R,M,O	4,8,11	S	CAEP	I	N	?
DE	Y	Y	Y(L)	2	R,M,W,O	1-8,11	C	CTBS**	E	N	N
FL	Y	Y	Y *****	1	R,M,W	3,5,8,10	C	SSATT	I	N	N

*Tested every 2 years

**Local Option

***New tests in 1985

****New tests this year

*****According to the state testing director the Florida assessment and competency tests are the same.

BEST COPY AVAILABLE

QUALITY INDICATORS SURVEY SUMMARY

General Characteristics

STATE	TESTING Have Program Y N	Used For: State Assessment Y N	Used For: Competency/ Proficiency Y N	No. of Testing Programs	ASSESSMENT PROGRAM: Areas Tested: Reading, Math, Other	Grade Levels	Selection Census Sample C S	Name of Test	Source of Items Internal External I E	Shared Items Y N	Major Planned Changes Y N
GA	Y	Y	Y	2	R,M,O	1-3,6, 8,10	C	SEVERAL,ITBS	I,E	N	Y
HI	Y	Y	Y	2	R,M,W	2,4,6, 8,10	C	SAT	I	N	N
ID	Y	N	Y	1	-	-	-	-	-	-	-
IL	Y	Y	N	1	R,M,W,O	4,8,11	S	-	I	N	Y
IN	Y*	N	Y*	1	-	-	-	-	-	-	-
IA	N	-	-	-	-	-	-	-	-	-	-
KS	Y	N	Y	1	-	-	-	-	-	-	-
KY	Y	Y	N	1	R,M,W,O	3,5,7, 10	C	CTBS	E	N	Y
LA	Y	Y	Y	2	R,M,W	7,10	C	-	I	Y	Y

*Program to start in 1985

BEST COPY AVAILABLE

QUALITY INDICATORS SURVEY SUMMARY

General Characteristics

STATE	TESTING Have Program	Used For: State Assessment	Used For: Competency/ Proficiency	No. of Testing Programs	ASSESSMENT PROGRAM: Areas Tested: Reading, Math, Other	Grade Levels	Selection Census Sample	Name of Test	Source of Items Internal External	Shared Items	Major Planned Changes
	Y N	Y N	Y N				C S		I E	Y N	Y N
ME	Y	Y	N	1	R,M,W,O	4,8,11	C	-	E	Y	Y
MD	Y	Y	Y	2	R,M,O	3,5,8	C	CAT	E	N	N
MA	Y	N	Y(L)	1	-	-	-	-	-	-	Y
MI	Y	Y*	Y	2	R,M,W,O	4,7,10	C	-	I	N	Y
MN	Y	Y	N	1	R,M,W,O	4,8,11	S	-	I	Y	Y
MS	Y	Y	Y**	2	R,M,O	4,6,8	C	CAT (old)	E	N	Y
MO	Y	Y	Y	2	R,M,O	6,12	S	-	I	N	Y
MT	Y	Y	N	1	R,M,O	6,11	V	-	I	N	Y
NE	Y	N	Y(L)	1		-	-	-	-	-	-

*According to the state testing director, the MI assessment and competency tests are the same.

**MS assessment program was dropped after 1983 and a new competency program is being implemented.

QUALITY INDICATORS SURVEY SUMMARY

General Characteristics

STATE	TESTING Have Program Y N	Used For: State Assessment Y N	Used For: Competency/ Proficiency Y N	No. of Testing Programs	ASSESSMENT PROGRAM: Areas Tested: Reading, Math, Other	Grade Levels	Selection Census Sample C S	Name of Test	Source of Items Internal External I E	Shared Items Y N	Major Planned Changes Y N
NV	Y	N	Y	1	-	-	-	-	-	-	-
NH	Y	N	Y(L)	1	-	-	-	-	-	-	-
NJ	Y	N	Y	1	-	-	-	-	-	-	-
NM	Y	Y	Y	2	R,M,O	3,5,8	C	CTBS	E	N	N
NY	Y	Y	Y	2	R,M,W	3,5,6	C	PEP	I	N	N
NC	Y	Y	Y	2	R,M,W,O	1-3,6,9	C	CAT	E	N	Y
ND	N	-	-	-	-	-	-	-	-	-	-
OH	Y	N	Y(L)	1	-	-	-	-	-	-	-
OK	N	-	-	-	-	-	-	-	-	-	-

BEST COPY AVAILABLE

QUALITY INDICATORS SURVEY SUMMARY

General Characteristics

STATE	TESTING Have Program Y N	Used For: State Assessment Y N	Used For: Competency/ Proficiency Y N	No. of Testing Programs	ASSESSMENT PROGRAM: Areas Tested: Reading, Math, Other	Grade Levels	Selection Census Sample C S	Name of Test	Source of Items Internal External I E	Shared Items Y N	Major Planned Changes Y N
OR	Y*	Y	N	1	R,M,W	4,7,11	S	-	I/E	Y	Y
PA	Y	Y	Y**	2	R,M,W,O	5,8,11	C	EQA	I	N	Y
RI	Y	Y	Y	2	R,M,O	4,6,8-10	S	ITBS	E	N	Y
SC	Y	Y	Y	2	R,M,O	4,7,10	C	CTBS-U	E	N	Y
SD	N	-	-	-	-	-	-	-	-	-	Y
TN	Y	Y**	Y	2	R,M,O	2,3,5-8 9-12	S	Metropolitan	E	N	Y
TX	Y	Y***	Y	1	R,M,W	3,5,9	C	-	I	N	Y
UT	Y	Y	N	1	R,M,O	5,11	S	CTBS-U	E	Y	Y
VT	Y	N	Y(L)	1	-	-	-	-	-	-	-

*Up to 1985 have tested only every 4 years, will be annual shortly in 1985 at grades 3,5,8,11

**New this year

***According to state testing director the TX assessment & competency tests are the same.

QUALITY INDICATORS SURVEY SUMMARY

General Characteristics

STATE	TESTING Have Program Y N	Used For: State Assessment Y N	Used For: Competency/ Proficiency Y N	No. of Testing Programs	ASSESSMENT PROGRAM: Areas Tested: Reading, Math, Other	Grade Levels	Selection Census Sample C S	Name of Test	Source of Items Internal External I E	Shared Items Y N	Major Planned Changes Y N
VA	Y	Y	Y	2	R,M,O	4,8,11	C	SRA	E	N	Y
WA	Y	Y	N	1	R,M,O	4,8,11	C/S*	CAT	E	N	Y
WV	Y	Y	N	1	R,M,O	3,6,9,11	C	CTBS-U	-	-	-
WI	Y	Y	Y**	2	R,M,W,O***	4,8,11	S	CTBS	E	N	Y
WY	Y	Y	Y(L)	2	R,M,W,O	3,4,7,8,11	S	NAEP	I	-	N

*Sample at grade 11

**Now in development

***Tested every three years

BEST COPY AVAILABLE

BEST COPY AVAILABLE

QUALITY INDICATORS SURVEY SUMMARY

General Characteristics continued

STATE	COMPETENCY PROGRAM: Areas Tested: Reading, Math, Other	Grade Levels	Selection Census Sample C S	Name of Test	Source of Items Internal External I E	Shared Items Y N	Major Planned Changes Y N
AL	R,M	3,6,9,11	C	ABCT,AHSGF	I	N	Y
AK	-	-	-	-	-	-	-
AZ	R,M,W	8,12	C	Local	I	N	Y
AR	R,M	3,4,6,8	C(4),S	CRT	I	N	Y
CA	R,M,O	?	C	Local	I		
CO	-	-	-	-	-	-	-
CT	R,M,W	4,6-8,9-12	C	CRT	I	N	Y
DE	R,M,W	8,11	C	Local	I		
FL *	R,M,W	3,5, 8,10	C	SSAT	I	N	N

*Same as assessment program.

General Characteristics continued

STATE	COMPETENCY PROGRAM: Areas Tested: Reading, Math, Other	Grade Levels	Selection Census Sample C S	Name of Test	Source of Items Internal External I E	Shared Items Y N	Major Planned Changes Y N
GA	R,M	4,8, 10-12	C	CRT	I	N	Y
HI	R,M,W	3,9-12	C	NWRL	I	N	Y
ID	R,M,W	8	C	-	I	N	Y
IL	-	-	-	-	-	-	-
IN	R,M,W,O	3,6, 8,10	C	NEW LOCAL	I	N	Y
IA	-	-	-	-	-	-	-
KS	R,M	2-4, 6,8,10	C	-	I	N	N
KY	-	-	-	-	-	-	-
LA	R,W,M	2-5	C	-	I/E	Y	Y

BEST COPY AVAILABLE

QUALITY INDICATORS SURVEY SUMMARY

General Characteristics continued

STATE	COMPETENCY PROGRAM: Areas Tested: Reading, Math, Other	Grade Levels	Selection Census Sample	Name of Test	Source of Items Internal External	Shared Items	Major Planned Changes
			C S		I E	Y N	Y N
ME	-	-	-	-	-	-	-
MD	R,M,W	7,9	C	-	I	N	N
MA	R,M	-	-	Local	-	-	-
MI *	R,M,O	4,7,10	C	-	I	N	Y
MN	-	-	-	-	-	-	-
MS	R,M,W (new)	3,5, 8,11	C	?	I	N	Y
MO	R,M,O	9-12	C	"Best Test"	I	N	N
MT	-	-	-	-	-	-	-
NE	R,M,W,O	5+	C	Local Option	I	?	?

*Same as assessment program.

General Characteristics continued

STATE	COMPETENCY PROGRAM: Areas Tested: Reading, Math, Other	Grade Levels	Selection Census Sample C S	Name of Test	Source of Items Internal External I E	Shared Items Y N	Major Planned Changes Y N
NV	R,M,	3,6, 9-12	C	CAT	I/E	Y	Y
NH	M,O	4,8,12	C,S	Local(new)	I	N	?
NJ	R,M,W	9-12	C	-	I	N	Y
NM	R,M,W	10	C,S	-	I	N	Y
NY	R,M,W	3, 10-12	C	Regents	I	N	N
NC	R,M,W*	1,2,3, 6,9,11	C	-	I	Y	Y
ND	-	-	-	-	-	-	-
OH	R,M,W	-	-	Local	-	-	Y
OK	-	-	-	-	-	-	Y

*New in 1985

QUALITY INDICATORS SURVEY SUMMARY

General Characteristics continued

STATE	COMPETENCY PROGRAM: Areas Tested: Reading, Math, Other	Grade Levels	Selection Census	Selection Sample	Name of Test	Source of Items Internal	Source of Items External	Shared Items		Major Planned Changes	
			C	S		I	E	Y	N	Y	N
OR	-	-	-		-	-		-		-	
PA	R,M	3,5,8	C		TELLS	I		Y		Y	
RI	R,M,O	8,10	C		Life Skills	I			N	Y	
SC	R,M,W	1-3,6, 8,11	C		Basic Skills Assessment	I			N	Y	
SD	-	-	-		-	-		-		-	
TN	R,M	9-12	C		-	I			N	Y	
TX*	R,M,W	1,3,5,7, 9,11,12	C		-	I		-		Y	
UT	-	-	-		-	-		-		-	
VT	R,M,W,O	1-12	-		Local	-		-		Y	

*Same as assessment program.

BEST COPY AVAILABLE

General Characteristics continued

STATE	COMPETENCY PROGRAM: Areas Tested: Reading, Math, Other	Grade Levels	Selection Census Sample C S	Name of Test	Source of Items Internal External I E	Shared Items Y N	Major Planned Changes Y N
VA	R,M,O	K-6, 10-12	C	?	I	N	Y
WA	-	-	-	-	-	-	-
WV	-	-	-	-	-	-	-
WI	R,M,O	3,7,10	C	Voluntary	I	N	Y
WY	R,M,W			Local			

BEST COPY AVAILABLE

APPENDIX 7

Common Test Linking Issues

Contents:

Notes by R. Darrell Bock
Letter to Leigh Burstein, from Robert L. Linn
Comments on Bock notes by J. Ward Keesling
Letter to Leigh Burstein, from Dale Carlson
Letter to Leigh Burstein, from Edward D. Roeber, Ph.D.
Letter to Leigh Burstein, from Tom Derins, Ed.D. and Jack Fyans, Ph.D
Letter to Leigh Burstein, from Lorrie A. Shepard

Using data from the National Assessment of Educational Progress to link state assessment results

R. Darrell Bock

University of Chicago

March 1, 1985

The National Assessment of Educational Progress (NAEP) as now conducted by Educational Testing Service can provide data that would enable assessment results from many of the states to be expressed on a common scale. Scaled in this way, these results would could be used in comparisons of educational attainment among the states participating in such an effort. Because of relatively small sample sizes in some states, present NAEP data can be used only for national and regional reporting, and not for between-state comparisons. In most states, however, the NAEP samples are large enough to support the scaling procedures required to establish a common basis for comparisons among state assessments.

The possibility of using the data in this way arises from NAEP's practice of assigning case numbers to each pupil's record in the public-use files. These case numbers are associated with corresponding pupil names and grades on rosters that are left in the possession of the schools where testing was carried out. (Pupil names never leave the schools.) For public schools at least, these rosters are presumably available to the state assessment programs, probably in the form of list prepared by the school that associates the NAEP case number with a corresponding state assessment case number.

A basis thus exists for identifying pupils who have taken both the NAEP assessment exercises and the state assessment exercises or tests, and for locating the item responses of these pupils on the NAEP public-use tapes. In those states, such as California, that test all pupils in the state at certain grade levels (i.e., perform a complete census of the state), these joint results will be available routinely when the NAEP and the state are testing at the same grade level. In states that test in only a sample of schools, special provision would have to be made to supplement the state sample with the schools in the NAEP sample.

That the NAEP testing is limited to grades 4, 8 and 11 will present difficulties, however, in those states that do not also test at these grade levels. Such states will have to arrange special administrations of their tests in the schools and at the grade levels of the NAEP testing. If, for example, a state system tests in sixth grade, that test would in most skill and content areas probably have sufficient range of difficulty to be successfully administered in the eighth grade for purposes of scaling. Even when there is no conflict of grade levels, differences in the time the tests are given during the school year may created a problem. If several months elapse between the NAEP and the state testing, special studies would have to be carried out to establish rate of change of scores during the year as a basis for correcting the results to a common date. But in all the problems, changes in the conduct of state assessment to conform to NAEP practices would be a better solution.

A more serious hindrance in the NAEP practice is their policy of testing only biennially, and only in a few content areas at one time. Thus, NAEP tested in Reading and Writing in 1983-84, and will test in Reading, Math, Science, and Computer Understanding in 1985-86, and in Reading, Writing, Math, and Science in 1987-88. Any attempt to link state assessment results using the NAEP data would therefore have to extend over a period of years, and even then might not include topics in state assessments outside these main areas. Nevertheless, the range of content in the complete NAEP cycle is broad enough to encompass the essential subject matter of primary and secondary schooling. Within the main areas, on the other hand, the NAEP exercise sets are large and varied and would probably parallel many exercises and items in the state assessments. Drawing these parallels in a comparable way in all of the participating state assessments would of course be essential to the linking of results. This problem is discussed below.

Another aspect of the NAEP design that creates difficulties for the analytical methods of scale linking is the sparseness of the present matrix sampling design. In the 1983-84 Reading assessment, for example, 139 items are matrix sampled in forms, each containing about 20 items. In any of the reading subareas sufficiently homogeneous to report in one score, any given pupil is presented only a small number of items, six to nine in most cases. As a result, any equating or scaling method that requires computation of scores for individual pupils will be impaired by the instability of scores computed from so few items. In particular, conventional linear or equipercentile equating of parallel forms, such as used in equating SAT forms, cannot be justified if, as is likely, the NAEP scores and the state assessment scores differ greatly in reliability.

Only those methods that estimate scaling constants directly from the item responses, without calculation of intervening scores, are suitable for this type of matrix sampled data. Fortunately, such methods are now available in item response theoretic (IRT) scaling based on marginal maximum likelihood procedures introduced by Bock and Aitkin (1981). These methods, which require large samples of persons but not large numbers of item responses from any given person, are ideally suited to the analysis of matrix sampled data. They are already used for that purpose by the California Assessment and for certain phases of the NAEP analyses.

The variant of these methods that would apply in the present case is a form of "old-test, new-test" technique. It is assumed that item parameters for the scale in question (the old test) have been estimated in the NAEP national sample. These item parameters are then used to compute the posterior distribution of the pupil's ability, conditional on his responding correctly or incorrectly to given items of the new test (comprised of items from the same content domain in the state test). In the Bock-Aitkin marginal maximum likelihood method of estimating item parameters, this distribution is represented by posterior densities on a finite number of points for purposes of numerical integration during marginalization. The item parameters of the new test are estimated by maximum likelihood from the sums these conditional densities over the sample of pupils (which is assumed to be large). The calculations are carried out iteratively by the so-called "EM algorithm" until stable values of the parameter estimates are obtained. These item parameter estimates are then used to compute scores for pupils in the state sample, preferably by the Expected A Posteriori (EAP) method (see Bock and Mislevy, 1982). The Posterior Standard Deviations (PSD) of these scores can be interpreted as standard errors for purposes of expressing their precision.

Provided the same prior distribution is assumed for purposes of marginalization (a normal distribution with mean 500 and standard deviation 100, for example), the EAP scale scores computed from the data of different states will have the same origin and unit and will thus be comparable for purposes of statistical comparisons between states.

Technically, this procedure is straightforward, computationally efficient, and statistically robust. The greatest difficulty in its implementation is the conceptual one of agreeing on common content domains in which the items from the participating state assessments should be classified for purposes of constructing attainment scales. The item domains must be essentially unidimensional, they must correspond to items in the National Assessment, and they must represent important areas of the curriculum. A common effort administered by the National Center for Education Statistics or the Education Commission for the States would obviously be required to obtain agreement on these points. An even better arrangement would be one involving NAEP in which the design of the national assessment is brought into better accommodation with the state assessments.

References

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 725-737.

Bock, R. D. & Mislevy, R. J. (1982) Adaptive EAP estimation of ability in a microcomputer environment. Journal of Applied Psychological Measurement, 6, 431-444.

University of Illinois
at Urbana-Champaign

Department of Educational Psychology

210 Education Building
1310 South Sixth Street
Champaign
Illinois 61820-6990

College of Education

217 333-2245

March 18, 1985

Professor Leigh Burstein
Department of Education
University of California, Los Angeles
Los Angeles, CA 90024

Dear Leigh:

I think that Darrell Bock's description of a procedure for using NAEP to link state assessment results is conceptually sound and technically feasible. What he has described is a viable means of obtaining much better comparative information than is currently possible provided states and key federal agencies have sufficient interest to cooperate. The main obstacles to successful implementation of the system revolve around content specification, grade-level coverage, timing of state assessments, and the need to collect and match state data for students in the NAEP sample.

Agreement on content domains and the classification of items from NAEP and each state assessment into those domains is crucial. The system cannot work without agreement. A carefully coordinated effort among interested states, key federal agencies, and NAEP would be needed to achieve the degree of consensus required for implementation and acceptance of the results. Your advisors who are directly involved with state assessments could give you a better idea of how feasible it is to accomplish this step.

As Darrell points out, the additional data collection would be required in states where the state assessment does not match NAEP in terms of grade levels covered or the time of year that data are collected. Resources obviously would need to be identified to cover the expenses of this additional data collection and analysis. Some cost estimates and maybe a pilot study in a couple of states would seem worthwhile.

It would also seem desirable to get a better idea about the extent of the mismatch problem. You may already have this from your review of state practices, but a comparison of content covered, grades included in the state assessments, and time of testing would be helpful. We would also need to have a sense of the viability of matching student data from NAEP with the state assessment results.

I think the idea has considerable merit. Perhaps the next step should be to see if any states have sufficient interest to pilot test the idea.

Best regards,

Robert L. Linn
Chairperson

RLL/jm

COMMENTS ON DARRELL BOCK'S TEST LINKING PROPOSALS

J. WARD KEESLING

1. Does Darrell have an idea of what numbers of items and students would be needed to make valid comparisons (or precise comparisons) among the states? Precision might be easy (?) to determine. Validity may be a more subjective call.

2. How many states would have enough students in G4, G8, or G11 with NAEP scores and state assessment scores to meet the criterion in #1?

3. How many states could be added if they would augment their samples with NAEP schools?

4. How many states could be added if a G6 state test could really do as well in G8 as a G8 test? How many items would have to come from the same "domains?"

5. How many states test at times not sufficiently close to NAEP tests?

6. Because most state assessments will include reading, and because the item types may be like those used in NAEP, this would be a good place to try a test case.

7. If at least 15 states can be found with reading assessments in the right grades at the right time of year, this would be a good test case. Data should be available from NAEP for 83-84 and 85-86.

8. In states planning to assess reading and/or math in 85-86, at about the same time-of-year as NAEP tests and in the same grades, start coordinating now to make it possible to try Darrell's idea.

9. Probably the most difficult part of this will be to identify items that truly belong in the same skill area or objective across the NAEP and SEA tests.

10. One could use the 83-84 data as a test case (probably only in reading, though).

11. A test case, such as this, may be the only way to examine the precision of state-by-state comparisons, and make projections about the numbers of people and items needed to make useful contrasts or rankings.

CALIFORNIA STATE DEPARTMENT OF EDUCATION
721 Capitol Mall
Sacramento, CA 95814

Bill Honig
Superintendent
of Public Instruction

April 2, 1985

Leigh Burstein
Department of Education
University of California
Los Angeles, California 90024

Dear Leigh:

Please forgive my tardy response to your letter of March 8. It has been more than a little hectic around here with the release of the grade 12 scores and the excitement surrounding the financial rewarding of schools for improving their scores under the "Cash for CAP" program.

I found Dr. Bock's summary of the proposed equating procedures consistent with our discussion last winter and as encouraging to me as when we first discussed them. My positive attitude rests on the moderately justifiable hope that we can have the best of both worlds--the manifold and manifest advantages of a "bottom up" approach to test content determination and credible state-by-state comparisons. Those comparisons will be harder to generate than those from a "national test" and will require some additional qualifications for interpretation, but the comparisons can be made.

Some states do not now test at the "right" grade levels or the "right" time of year. The two-choice solution to these problems is totally compatible with the American philosophy of federal-state relations: (1) the remaining states will join the NAEP pattern, or (2) the NAEP grade levels, although selected on solid grounds, will be judged not to meet the needs of most states and districts and, therefore, ought to be changed. (A one-time break in the longitudinal comparisons could be accommodated by NAEP with no substantial increases in testing time for that one year.)

Similarly, NAEP's biannual assessment schedule does not seem insuperable. It means that new state tests could be calibrated, without additional testing, only every other year. It would be nice to have annual state-national comparisons, but most of the states could still be compared on an annual basis.

We are fortunate that Dr. Bock has developed such innovative and powerful procedures to handle what would otherwise be in intractable problem (i.e., the sparseness of the NAEP sampling design), thereby avoiding a complete redesign of NAEP's procedures. I hope that Dr. Bock's procedures can be put to the test under these circumstances, which are just different enough from the California application to make them challenging.

A critical issue, of course, is that of test content. Is there sufficient agreement among the states on the most important content to be tested? I think so. The fact that the content focus is always changing complicates the process because the changes are not uniform across the states. But that is a small price to pay for the assurance of a timely and genuine content validity as it reflects the consensus of local concerns. I am looking forward to hearing more of the progress you are making in probing these issues during the pilot study.

To summarize, I think we are on the right track. I am biased, I admit. This "bottom up" approach to gaining agreement on test content is consistent with our efforts to design a comprehensive assessment system in California--one that provides the public with valid comparative information reflecting core content, yet allows school districts to assess other objectives in sufficient scope and depth to meet their local needs.

I hope your surveying and summarizing are going well. I am looking forward to seeing the results of your efforts later this spring.

Sincerely,

Dale

Dale Carlson, Director
California Assessment Program
(916) 322-2200

DC:cc

STATE OF MICHIGAN

DEPARTMENT OF EDUCATION

Lansing, Michigan 48909

PHILLIP E. RUNKEL
Superintendent
of Public Instruction

STATE BOARD OF EDUCATION
NORMAN OTTO STOCKMEYER, SR.
President
BARBARA DUMOUCHELLE
Vice President
BARBARA ROBERTS MASON
Secretary
DOROTHY BEARDMORE
Treasurer
DR. EDMUND F. VANDETTE
NASBE Delegate
CARROLL M. HUTTON
CHERRY JACOBUS
ANNETTA MILLER

GOV. JAMES J. BLANCHARD
Ex-Officio

April 10, 1985

Dr. Leigh Burstein
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90224

Dear Leigh:

As you requested, I am providing you my comments and reactions to the paper by Darrell Bock that you sent me. I am sorry that I will be unable to join you in Washington, D.C., April 15th and 16th, but I have a conflict with a meeting of our State Board of Education on those dates. My reactions to Darell's ideas for using NAEP to link state assessment results are based both on my experience of directing the program here in Michigan, as well as having been a NAEP staff member in the late 60's and early 70's. Therefore, I am familiar with NAEP, its objective and item development procedures and sampling design.

NAEP has proposed a direct state-NAEP comparison for each state (which if all states elected would allow state-to-state comparisons as well). I am opposed to it for Michigan because 1) the skills tested don't match Michigan objectives; 2) the skills were by and large developed without the input of state departments of education curriculum specialists; 3) the range of diffuculties of items NAEP uses is purposely manipulated to produce a test with one-third very difficult (= .1) items, one-third medium difficult (p = .5) items and one-third easy (p = .9) items. In Michigan (and many other states), what is tested is what all students should know, regardless of the distribution of difficult or easy items; and 4) the cost of a state sample on NAEP is greater than or equal to that of testing all students at several grades in one subject area. Every-pupil data is far superior to sample data for improving schools. Since we are trying to add another subject area to the every-pupil assessment program here, cost is a very big item.

I was hoping, when I proposed to use NAEP as an anchor test, that little additional NAEP-type testing would be needed. However, Darrell states on page one of this paper that additional testing would be needed in states that only test in a sample of schools, which do not test students in grades 4, 8 and 11 or which test at a different time than NAEP's planned "spring" testing period of March-May. While Michigan tests all students, we test early September to early October in grades 4, 7 and 10. At least special bridge studies would be needed and perhaps it would be necessary to test samples of students in grades 4, 8 and 11 in the spring each time NAEP tests are given.

However, I do not see that "changes in the conduct of state assessment to conform to NAEP practices would be a better solution." I have cited the lack of conformance of skills tested, how the tests are built (NAEP never has specified that students ought to know anything they test), plus the very high cost of NAEP for just sample results. NAEP simply has limited utility in states that have strong state assessment programs. Since NAEP's purposes and methodology are different, it doesn't make sense to impose it on states.

On the other hand, there are quite a few similarities among the states with strong assessment programs. It would make more sense to capitalize on the commonalities of these programs and impose it back on NAEP. NAEP could collect, as one part of its data collection efforts, how the nation's students are doing on the skills that states think are most critical for all students to know in mathematics, reading and other areas. I believe the CCSSO proposal to develop a common core of competencies is heading in this direction, although I don't believe that they make any mention of using NAEP to collect the data.

While I understand that NAEP could be used to link state assessment results, my feeling is that it isn't worth the costs, either financially or curricular. I believe that whatever measure is used to compare the schools in Michigan with those of other states should first be defensible on the basis of content. I fear that if NAEP is used to link states and considerably more testing is needed, the foucs will be on NAEP performance, not state assessment results. I cannot defend the NAEP objectives as appropriate for all students here. Since the development of an adequate linking measure will take time, I believe we should direct our efforts to more curricularly-defensible techniques, such as the CCSSO proposal.

I hope these comments will prove useful to you and the committee. If you wish for me to elaborate on any of the points I have made, please feel free to contact me. I am sorry that my schedule won't permit me to join you next week.

Sincerely,

Ed

Edward D. Roeber, Ph.D.
Supervisor
Michigan Educational
Assessment Program

EDR/pg

Illinois
State Board of
Education

EDUCATION IS EVERYONE'S FUTURE

100 North First Street
Springfield, Illinois 62777
217/782-4321

Walter W. Naumer, Jr., Chairman
Illinois State Board of Education

Ted Sanders
State Superintendent of Education

April 11, 1985

Leigh Burstein, Ph.D.
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

Dear Leigh:

We appreciated the opportunity to review the proposal by Darrell Bock which came with your March 8, 1985 correspondence.

There are several questions which are raised in the issues discussed by Bock. These are:

1) Which prior distributions will be chosen to generate the posterior densities in this model? Should the priors be baseline information from past NAEP assessments? Should the prior vary state to state or be set nationally? Furthermore, who should have the responsibility to decide what these priors should be?

2) It is true that posterior density estimates of scores can be generated by the model presented by Bock. A lingering question is how well will scores produced by such a model represent the students from which they are derived? That is, how will the psychometric model presented by Bock interweave with a sampling model to produce results proportionate to the number and type of students spread out across the United States? Would the posterior score estimates by Bock then be weighted by sampling parameters to produce results to each state which would be useful to and representative of that state.

3) A related issue is that of the size of the population needed for this numerical integration. It would appear that the requisite sample size for such integration and maximum likelihood estimates would be large. As the number of educational domains and items therein increase, the N required will also increase. The need for certain levels of total N for psychometric stability may militate against the needs for certain representative N by states discussed in (2) above.

Chicago Office
188 West Randolph
Chicago, Illinois 60601
312/793-2220

Southern Illinois Regional Office
First Bank and Trust Building
Suite 214, 123 South 10th Street
Mt. Vernon, Illinois 62864
618/242-1676

An Equal Opportunity/Affirmative Action Employer

4) One major concern is that of dimensionality. Will the item response analysis find unidimensionality (even with one domain) across the items and many different types of students from throughout the United States? A major effort could be conducted on a state by state basis (of those states participating) to assure the relevance of the items used with the curriculum taught in the state. It is simply not sufficient to have NAEP define "important areas of the curriculum." Work by Harnisch and others have shown how the measurement models vary by curricular differences among schools.

5) One suggestion might be the adoption of a weighted collateral information model of the sort discussed by Novick and Jackson (1974). That is, the data used for comparisons among states and for students would be a weighted composite of several components tapping the different leveTs in this analysis. Each component weighted by its own generalizability co-efficient. That is, the students' score would be weighted by the generalizability of data at student level added to the state means weighted by generalizability of data from that state, and combined with the overall national mean weighted generalizability at the national level. We have attached an article which describes this process.

6. On what basis can a claim be made that the NAEP tests "probably have sufficient range of difficulty?" We have not seen such empirical evidence. In Illinois, scaling of NAEP items by Logist V have shown them to be restricted in their difficulty, usually to unacceptably low levels. For example, the parameters of the NAEP items were much lower in difficulty and discrimination than those designed by our own staff and committees. In reading, for example, the NAEP items were answered correctly by 80% to 90% of our students.

Cordially,

Tom Kerins

Tom Kerins, Ed.D.

Jack Fyans

Jack Fyans, Ph.D.
Department of Planning,
Research and Evaluation

LMP4698f

July 26, 1985

Dr. Leigh Berstein
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, CA 90024

Dear Leigh:

I offer this letter as a minority report in contrast to your conclusions from the State Assessment/Quality Indicators Project reflected in your letter to Emerson Elliott on 22 April 1985. I believe that the State Assessment Consortium option which you are advocating is by far the most costly and potentially the most intrusive in terms of local testing demands, despite state ownership. Let me spell out what I believe are the detractions to the State Assessment Consortium option. Then, I will consider the Standardized Tests model which is the most cost effective for certain limited purposes. Finally, I will argue for the feasibility of an "Expanded NAEP" in contrast to the equated State Assessments model.

Obviously, the relative strengths and weaknesses of these options depend on the purpose of the assessment. Is the primary audience to be policy makers at the federal level, who seek valid state-by-state comparisons of pupil learning? Must the needs of state-level policy makers also be addressed? If so, will state-level decision makers be content with a summary report card comparing their state to other states and to the nation? Or, will they require more detailed, "instructionally diagnostic," information about relative strengths and weaknesses within broad subject areas? The latter, of course, requires a more sensitive measurement instrument with concomitant increases in cost. Also, note that this latter type of comprehensive in-depth assessment is not in keeping with the usual connotations of the term "indicator."

STATE ASSESSMENT CONSORTIUM

I did not raise any technical objections to Darrell Bock's memo of March 1, 1985, describing the procedure for linking state assessments via NAEP. Dr. Bock was very accurate in anticipating the number of special samples and special studies that would be required to implement such a design. It was not his purpose to offer a cost analysis. (However, once one attaches reasonable numbers to each special provision, the cost implications are clear.) Committee members who favor this plan obviously value state ownership of the test content so highly that they believe the extra cost in warranted.

COST. If EVERY state gave tests in the SAME CONTENT AREAS as NAEP, at the SAME GRADE LEVELS as NAEP, at the SAME TIME OF YEAR as NAEP, in precisely the SAME SAMPLE OF SCHOOLS as NAEP, and if the NAEP SAMPLES WERE ALWAYS LARGE ENOUGH, the linking of state assessments would clearly be cheaper than an expanded National Assessment because the extra cost of the equating analysis would more than be off-set by the savings in test administration, i.e., no additional sampling or testing would be required. However, none of these ideal matches are satisfied, hence, the need for expensive corrective strategies.

If one wishes to have data for all 50 states, which is presumably essential for FEDERAL audiences, then the equivalent of an expanded NAEP is needed in those states without a state assessment AND in those states for whom current NAEP samples are too small. According to your survey, at least 12 states do not have ANY state assessment or minimum competency tests. (I have excluded local district tests since these would require equating or anchor studies district-by-district.) Many more states are missing tests at one or more of the NAEP grade levels OR can be expected to have too sparse a NAEP sample for equating purposes. Because NAEP selects a sample to be representative of an entire region the state samples are not necessarily large enough EVEN FOR EQUATING as Darrell pointed out. Smaller population states such as New Mexico, Nevada, Maine, Montana, Alaska, would likely require augmented NAEP samples. Thus, in any kind of cost comparison the cost for these states would be roughly comparable to the expanded NAEP design.

Most states with testing programs test in reading and math and usually at at least two of the three school levels, elementary, middle, and high school. As Darrell has indicated, whenever a state does not test at grades 4, 8, and 11, the state will have to arrange special administrations of their tests at NAEP schools and at NAEP grade levels. Although I am willing to acknowledge that equating samples do not have to be as large as assessment samples, I am assuming that in these cases of mismatched grades it would not be possible to use the DATA from the regular state assessment only the TESTS. If the data from the next higher or next lower grade were used, it would require a statistical extrapolation of performance level that I do not believe is defensible politically. If you are willing to live with such extrapolations, because they provide rough "indicators" of the relative standing of state systems, fine; but then I don't think you should be so snobbish about nuances of content quality. Of course, if you don't extrapolate from the regular state assessments, then the NAEP-grade special administrations must be large enough to stand as the assessment samples.

A few more states, Connecticut, Illinois, Minnesota, Missouri, Oregon, Rhode Island, Tennessee, Utah, and Wisconsin, will require special state sampling if they do not already have a "piggyback" arrangement with NAEP. These states test only a sample of pupils rather than every pupil at a grade level. Unless there has been a specific contract with NAEP (which was at one time true in Minnesota, I know) the NAEP sample is not likely to coincide with the state sample. Thus, the state will have to add NAEP schools to the state sample.

Whenever state tests are not given at the same time of year as NAEP, special studies will have to be carried out to adjust performance to a common time. Now that NAEP is moving to a spring testing period (February - May), I expect this will be the least frequent source of difficulty. When they do occur, of course, these studies are an additional expense.

If the State Assessment model is put forward as the preferred solution, it should be accompanied by a realistic cost analysis.

INTRUSION. The equating plan relies heavily on the cooperation of local school personnel (the principal and secretary in each building). Retrieving names associated with NAEP IDs can be done and ETS has had reasonable success doing so in small-scale studies of their own. An equivalent effort is required to match names to state IDs. Even if we are only speaking about a day of the secretary's time, and even if a battalion of field supervisors are hired ($$$) to see it done properly, I believe there will be errors and missing data created by the negative reaction. This is an unforeseen burden falling on those who agreed to be NAEP schools.

Even more intrusive is the implicit expectation that ultimately the costs of such a system will diminish as the STATES ADJUST THEIR ASSESSMENTS TO THE NAEP DESIGN. (Dale Carlson mentioned in his memo that NAEP might also change to fit more popular grade levels. But, when you consider that the precise choice of grades is arbitrary and that there is no other more prevalent set of grades than 4, 8, 11, the direction of conformity is clear.) It is ironic that a plan that has state ownership as its prinicpal attraction would have such compliance as its goal. Not only would states disrupt their own change data but then there really would be only one all-powerful federalist system. If you didn't like what this test said about you, there would be no other recourse. Whereas, a NAEP test would never be so potent, especially if on a different day the headlines were about the state test and progress over time.

NOT ALL STATES. Of course, my Cassandra-like cost analysis is exaggerated if you have no intention of including all 50 states. If, instead, you included only 25 states who were interested, had large populations and their own extensive assessment programs, and fit the NAEP design at least in part, then the cost TO THEM would not be as great as the cost of an expanded NAEP to the federal government. Let us be clear, however, that such a plan would only serve state-level policy makers by providing them with national comparisons. In which case, it is not apparent to me why we are addressing such advice to Washington officials unless they see themselves as facilitators of state-level decision making.

I do not believe that the documents circulated thus far have spelled out for all to see that the State Assessment model is a not-all-states solution.

"BOTTOM UP CONTROL OF CONTENT." There is a troubling contradiction in believing that individual state tests are importantly different enough to justify the elaborate linking design but similar enough to satisfy the requirements of IRT. I have seen laughable applications of IRT calibration where the limited number of items per subtest (4-5) was overcome by a total-test analysis (assuming unidimensionality) but then users expected to obtain differential diagnostic information from the subtest. Dr. Bock has never been guilty of such foolishness.

Instead, he has advocated scaling of "indivisible curricular elements." Less sophisticated audiences are more likely to trust in the magic of IRT and believe that they can have their cake and eat it, too.

Let's make it explicit that if a state has a unique content element that is not represented in the NAEP test, it cannot be equated. In essence, the grand scheme allows states to be ranked on their own items that most resemble the NAEP content. It is a fiction that their unique objectives can raise them on the NAEP ranking.

READING, A SPECIAL CASE. Finally, the enthusiasm for the State Assessment model should be tempered by the warning that the equating strategy could work in reading and NOT in other subjects. Reading is not only the most universally assessed area, it is also the most uniformly defined and best satisfies the unidimensionality requirements.

STANDARDIZE TESTS. Nearly every school district in the country administers standardized tests of achievement. Only about five or six major batteries account for 90% of the market. One way to gather credible comparative data is to draw a representative sample of school districts in each state and to require (presuming a federal mandate) selected districts to report their aggregate scores by grade tested, sample size, time of testing, and form of the test used. Normative standing for each district and then state could be averaged across grades and tests based on equivalencies derived from one national anchor study. Unlike the State Assessment model, separate equating studies would not have to be done in each state. Because the districts would supply the data and the anchor study would supply the conversion metrics, cooperation from the best publishers would not be essential.

I would never advocate such a plan as a comprehensive in-depth assessment. But, if what you want is an "indicator" of relative state achievement, then it would be the cheapest but adequate model. The logistics of DISTRICT data collections would be more feasible than the pupil-level coding of the state linking design. Furthermore, it would be easy to collect demographic indicators at the same time. Any of these plans must make provision for assessing background factors (e.g., mobility, percent below poverty) against which achievement results are interpreted.

EXPANDED NAEP. An "expanded National Assessment" would involve increasing the current NAEP samples in most states to permit state-level results. If you believe the tests are narrow instruments or not as good as some state tests then the content could also be expanded either by lobbying NAEP or by making agreements with a few states to share their items. (If you really believe the NAEP tests are so bad, you should be lobbying ETS anyway.) The expanded NAEP model would be cheaper than an all-50-state implementation of the State Assessment model. The most accurate cost estimates can be obtained for this design because the cost is directly tied to sample size and because ETS has already had experience with piggybacking and with the southern consortium.

As I mentioned then, the two objections to the NAEP solution are (1) the limitation of the tests, and (2) the political undoing of NAEP by making it a national test with authority.

I believe you are being overly esoteric in criticizing the NAEP tests. Equally distinguished groups of subject matter experts were convened to create those tests as those in the respective states. And, as I indicated above, if your criticisms are warranted, the right thing to do is change NAEP. In fact, however, I believe that only a few states can boast tests that are "better" (in terms of content coverage or item quality, not just better suited to their own needs) than the NAEP tests. Because of the matrix design, in fact, NAEP content domains are much more comprehensively assessed than in most state tests. Are you concerned that they don't test higher order cognitive skills? If you're right, these elements would be missing from the equating design, as well.

If you are worried about NAEP's political future, consider that with the move to ETS, NAEP has already abandoned its character as the dull monitor of an achievement time series. The NAEP staff have promised to deliver a national report card and are aggressively trying to make the NAEP data as visible and useful (hence political) as possible. Furthermore, your state assessment model with its dependence on NAEP and its evolutionary adaptation to the NAEP design will eventually give the NAEP tests the authority you seek to avoid. The dozen biggest states might be likely to keep their own assessments, but if the State Assessment model were fully in place, one wonders if smaller states would be motivated to maintain their own assessments instead of adopting the NAEP tests as well as the NAEP schedule. When you come right down to it, it is the largest states with visible assessment programs for whom the ownership issues are the most salient. Smaller states might prefer the NAEP design to the expensive linking system.

Please find an appendix somewhere for my contrary opinions.

Sincerely,

Lorrie A. Shepard

Lorrie A. Shepard
Professor

sm

APPENDIX B

7/1/86

SUMMARY OF DOCUMENTS PROVIDED BY STATE TESTING PROGRAMS
STATE TESTS AS QUALITY INDICATORS PROJECT

STATE	CODE*	REPORT TITLES	YEAR
Alabama	St	High School Graduation Examination State Report: Reading	1983
	St	Basic Competency Testing Program State Report: Reading (Grade 3)	1983
	St	Chief State School Officer Summary Report: California Achievement Test, 1977 Edition	1983
Alaska	C	Alaska Statewide Student Assessment Program. Reading Skills Objectives: Grade 8. Field Review Edition	Jan. 78
	C/T	Portland Developmental Items. Mathematics: Grades 4 & 8. Reading: Grades 4 & 8	1982
	T	Statewide Achievement Test in Reading and Mathematics: Grade 4	--
	St	Report on the 1981 Alaska Statewide Assessment Tests	1981
	St	Alaska Statewide Student Assessment: A Comparison of the 1977 and the 1979 Assessment Results	--
	St	Results of the 1983 Statewide Assessment Tests	1983
	M	Alaska Instructional Diagnostic System: Pilot Test Results	1978
	M	AIDS - An Evaluation of the Use of AIDS by Teachers	--
	TM	AIDS - Skill Sheets Reading (General Information)	--
	C/TM	Structural Analysis (Skill Survey Sheets, Reading)	--
	TM	AIDS - Lower Level Skill Surveys (General Information)	--
	M	AIDS - Overview	--
	TM	AIDS - Upper Level Skill Surveys (General Instructions)	--
	M	AIDS - Workshop Overview	--
	T	AIDS - Student Booklet(Mathematics). Upper Level Skill Surveys	1977
	C	Cross-Reference Guide (Computational Skills & Alaska objectives and Items Bank	--

*Key to document attached at end.

BEST COPY AVAILABLE

STATE	CODE	REPORT TITLES	YEAR
Arizona	St	Arizona Pupil Achievement Testing: Statewide Report	1984
Arkansas	St/D	Analysis & Interpretation of the Results of the Arkansas Norm-Referenced Testing Program	1983-84
	St	Analysis & Interpretation of the results of the Arkansas Minimum Performance Testing Program	1983-84
California	D	California Assessment Program: Four Year District Summary	1983-84
	C	Survey of Basic Skills: Grades 3 & 6. Rationale & Content	1983-84
	C	Survey of Academic Skills: Grade 8. Skill Areas Assessed in Reading & Written Expression. Rationale & Content	March 1984
	C	Survey of Academic Skills: Grade 8. Skill Areas Assessed in Mathematics. Rationale & Content	March 1984
	St/C/Tc/D	Survey of Basic Skills: Grade 6 Part I: Content Area Summary Part II: Program Diagnostic Displays Part III: Subgroup Results Part IV: Using Survey Results Part V: Interpretive Supplement and Conversion Tables	1982
	St	Student Achievement in California Schools: Annual Report	1982-83
	P/D	Profiles of School District Performance. A Guide to Interpretation	1982-83
	C	Test Content Specifications for the Survey of Basic Skills: Written Expression and Spelling, Grades 6 & 12	1975
	Tc/Su	Interpretive Supplement to the Report on the Survey of Basic Skills: Grade 6	1980
	St	Student Achievement in California Schools: Annual Report	1981-82
	P/D	Profiles of School District Performance	1979-80
	C	Test Content Specifications for the Survey of Basic Skills: Mathematics, Grades 6 & 12	1975
	C	Survey of Basic Skills: Grade 12	1981
	Su	Survey of Basic Skills: Interpreting Results, Grade 12	1984
	C	Test Content Specifications for California State Reading Tests: Grades 2,3,6,12	1975

STATE	CODE	REPORT TITLES	YEAR
Connecticut	M	How Testing is Changing Education in Conn.	1983-84
	P	Mater & Remedial Standards for the 4th Grade	1984
	St	Conn. Assessment of Education Progress	1983-84
	P	Presentation on Conn. Assessment of Education Progress (CAEP) Program Update	1984
	M	CAEP IV Grade 8 Objectives	1983
	M	Teaching Thinking and Problem Solving	1985
	St	Conn. Assessment of Educational Progress, Social Studies, Overview of the Assessments	1982-83
	Tc	Conn. Assessment of Eductional Progress Summary & Interpretations	1982-83
	St	Business & Office Education Brochure, Overview of the Assessment	1983-84
	St	Social Studices Summary & Interpretations	1982-83
	St	Art & Music, Summary & Interpretations	1982-83
	St	Sience, Summary & Interpretations	1979-80
	St	Math, Gr. 11, Summary & Interpretations	1979-80
	T	Conn. Basic Skills Proficiency Test, Math, Form B	1982
	T	Conn. Basic Writing Skills in Language Arts, Form B	1982
	T	Mathematics, Gr. 11, 8, 4	1979-80
	St	Conn. Ninth-Grade Proficiency Test, Summary Report	1980-81
	St	Conn. Basic Skills Proficiency Test Results	1984-85
	M	Objectives and Standards for Testing Program	1985
	M	How Testing is Changing in Conn. (Article)	1985

BEST COPY AVAILABLE

STATE	CODE	REPORT TITLES	YEAR
Delaware	St	Educational Assessment Program. Statewide Test Results: Summary Report	1983-84
	P	Delaware Educational Assessment Program: Profile Report	March 1984
	I	Delaware Educational Assessment Program: Individual Item Report	March 1984
	G	Group Right Response Report	March 1984
	St	Delaware Educational Assessment Program: Statewide Testing Results	1983
Florida	St	SSAT One Results. Student Assessment Test, Part I. Grades 3,5,8	October 1983
	C	Item Specifications for the State Student Assessment Test, Basic Skills	1985
	St	SSAT One & Two Results. Grades 3,5,8,10	1982-83
	C	Minimum Student Performance. Standards for Florida Schools. Grades 3,5,8,11 (Reading, Writing, Mathematics)	1985
	St	State, District, & Regional Report of Statewide Assessment Results	1983
	Tc	Technical Report	1983-84
	Su	Statistical Supplement to the Technical Report	1983-84
Georgia	C/Tc	First, Fourth, and Eighth Grade Criterion-Referenced Test: Objectives and Assessment Characteristics	1983
	T/C/St	Student Assessment: Criterion-Referenced Tests and Basic Skills Tests (Content and Results)	1983-84
	C/Tc	Criterion-Referenced Tests (Mathematics and Reading Tests): Objectives and Assessment Characteristics for Third and Sixth Grade	1984 1984

STATE	CODE	REPORT TITLES	YEAR
Hawaii	IN	Teacher's Handbook on Essential Competencies (Draft)	1983
	St	Summary Report of Statewide Testing Program	1984
	St	Summary Report of Statewide Testing Program	1983
	M	Graduation Requirements and the HI State Test of Essential Competencies (HSTEC), effective 1983	1982
Idaho	T	Idaho Proficiency Test: Mathematics, Reading, Spelling	--
	C/M	Proficiency Testing Program	1981
	Su	Interpretive Guide to Computer Printouts	1982
	TM	Test Administration Manuel	1984
	St	Report on Idaho Proficiency Test Results	1983-84
Illinois	St	Summary of the 1982 Mathematics Results of the Illinois Inventory of Educational Progress	1982
	St	Student achievement in Illinois: An Analysis of Student Progress	1982-83
	St	School District Organization in Illinois	1985
	T	The Illinois Inventory of Educational Progress: Grades 4,8,11	1982-83
	T	The Illinois Inventory of Educational Progress: Grades 4,8,11	1985
	C	Curricular Analysis of the 1982 Mathematics Results of the Illinois Inventory of Educational Progress	1982
	D	Student Achievement in IL: An Analysis of Student Progress	1985
Indiana	Su	Design Specifications, the law, draft of questions/answers, and other related papers.	(began Feb. 85)

BEST COPY AVAILABLE

STATE	CODE	REPORT TITLES	YEAR
Kansas	St	Kansas Minimum Competency Testing Program Report (Rating Scales)	1983
	St	Report of Research Findings: The Kansas Competency Testing Program	1980
	C	Kansas Minimum Competency Objectives	1984-85
	Sc/D	Identifying Minimum Skills	--
	St	Kansas Minimum Competency Assessment Report: Reading and Mathematics	1982-83
Kentucky	St	Comprehensive Tests of Basic Skills. Statewide Testing Results: Grades 3,5,7,10	Spring 1984
Louisiana	C	Louisiana Basic Skills Testing Program. Language Arts & Mathematics Item Specifications: Grade 2 (1981), Grade 5 (1984-85), Grade 4 (1983-84), Grade 3 (1982-1983)	
	T/TM	Louisiana Basic Skills Testing Program. School Test Coordinators Manual: Grades 2, 3, 4, & 5 Basic Skills Tests	1984-85
	St	Basic Skills Testing Program. Annual Report: Grades 2,3,4	1983-84
Maine	St	Assessment of Educational Progress: Reading & Language Arts Results. Grades 4,8,11	1982
	St	Maine Assessment of Educational Progress: Reading & Language Arts. Summary & Interpretive Report, Grades 4,8,11	1982
	Tc	Maine Assessment of Educational Progress: Reading & Language Arts. Technical Report, Grades 4,8,11	1982
Maryland	St	Facts About Maryland Public Education. A Statistical	1983-84
	Tc	Facts About the California Achievement Test, Maryland Functional Reading Test, & Maryland Mathematics Test.	1980-84
	St	California Achievement Test Results: Grades 3,5,8. Maryland Functional Reading Test: Grades 9-12. Maryland Functional Mathematics Test: Grade 9.	1980-84
	IN	Project Basic Instructional Guide: Volumes V & VI. Functional Mathematics & Functional Reading	1981-82
	St	Maryland Accountability Testing Program: Annual Report	1981-82 1982-83

BEST COPY AVAILABLE

STATE	CODE	REPORT TITLES	YEAR
Massachusetts	Tc	Basic Skills Improvement Policy. An Implementation Evaluation of the Basic Skills Improvement Policy: Technical Appendix.	1983
	St	Basic Skill Improvement Policy. Statewide Summary of Student Achievement of Minimum Standards in the Basic Skills of Reading, Writing, & Mathematics	1981, 1983, 1984
Michigan	St	Mathematics Education Interpretive Reports: Grades 4,7,10	1980-81
	Su	MEAP Support Materials for Mathematics	
	C	Minimal Performance Objectives for Mathematics & Communication Skills (Reading, Writing, Speaking/Listening)	
	C/Tc/Su	MEAP Handbook	1984-85
Michigan (Cont.)	Tm	Coordination & Administration Manuel: Grades 4, 7, 10	1984
	St	MEAP Statewide Results	1983-84
	Tc	Technical Report: Volume I & II	1980-81
Minnesota	T	Minnesota Statewide Educational Assessment in Art	1981-82
	Tc/St/D	Performance in Basic Mathematics	1979-80
	T	Minnesota Statewide Educational Assessment in Literature and Mathematics	1982-83
	St	Results of a Statewide Assessment Program Utilizing the Minnesota Secondary Reading Inventories	1982-83
	St	Results of Minnesota Statewide Educational Assessment in Music	1980-81
	T	Statewide Educational Assessment in Reading	1981-82
	St	Results of Statewide Educational Assessment in Social Studies	1981-82
Mississippi	Su	Programs on Performance Testing Accepted October 18, 1984. Literature Regarding this & Preliminary Facts Pertaining to the 11th Grade Tests.	
Missouri	St	Statewide Assessment Data Summary (Grade 12)	Fall 1983
	Su	Interpretive Report: Grade 12 and 6	1976-77
	Su	Educational Goals	--
	C	Educational Objectives	1982

BEST COPY AVAILABLE

STATE	CODE	REPORT TITLES	YEAR
Missouri (Cont.)	C	Performance Indicators for Educational Objectives: Grades 6 and 12	1974-75
	St	State Assessment Data Summary	April
Montana	T	Montana School Testing Service Test Booklet: Grade 6 & 11	1984
	St	Results of 1984 Montana School Testing Service for Montana State Totals (Elementary & Secondary)	1984
Nevada	St	The Nevada Proficiency Examination Program. A Brief Description and the Results of the 1983 Examinations	1983
New Hampshire	St	Summary Report on Educational Assessment Program	1978,80
New Jersey	St	Statewide Testing System (New Jersey Public Schools)	Jan. 83
	St	Minimum Basic Skills Test Results	1983-84
	T	High School Proficiency Test: Grade 9. Statewide Results	1983-84
	C	Statewide Testing System. High School Proficiency Test. Directory of Test Specifications & Items	1983-84
	TM	Statewide Testing System. High School Proficiency Test. School District Guidelines	1983-84
	T	Minimum Basic Skills Test: Grade 9	1983-84
	TM/Su	Minimum Basic Skills Test: School District Guidelines	1983-84
New Mexico	St	Highlights of Results: High School Proficiency Examination	Spring 1984
	D	School District Profile	1982-83
	St	ACT & SAT Results	1982-83
	St	Standardized Testing Program Report	1982-83
	Su	Dropout Study	1982-83
New York	Su	Regents Competency Testing Program (Information Bulletin)	1982
	TM	Regents Examinations & Competency Tests. School Administrator's Manuel	1983
	TM	Reading Test: Grades 3 & 6. Manuel for Administrators & Teachers	1984

BEST COPY AVAILABLE

STATE	CODE	REPORT TITLES	YEAR
New York (Cont.)	TM	Mathematics Tests: Grades 3 & 6. Manuel for Administrators & Teachers	1984
	TM	New York State Preliminary Competency Test in Reading. Manuel for Teachers & Administrators	1982
	T	Writing Test: Grade 5	1984
	T	Preliminary Competency Test in Writing	1984
	TM	New York State Pupil Evaluation Program & Preliminary Competency Tests	1984-85
	St	Grade 3 & 6 Reading and Math Test Results	1983
	Sc/D	Regents Examination, Competency Test, & High School Graduation Statistics	1982-83
North Carolina	St	Competency Test Program: Report of Student Performance	Fall 1983
	St	Annual Testing Program: Basic Skills. Report of Student Performance Update from Spring 1981 to Spring 1984	Spring 1981-84
Oregon	St	Oregon Statewide Assessment: Summary Report	1982
Pennsylvania	Tc	An Analysis of Changes Across Time for Schools Participating in Educational Quality Assessment	1978-81 1979-81
	I/C/St	Educational Quality Assessment(EQA). Results from 1978-1981 Grades 5, 8, & 11	1982
	I/C	Getting Inside the EQA Inventory: Grades 5, 8, & 11	1982
	TM	Testing for Essential Learning & Literacy Skills(TELLS): Guidelines for Testing	1984
	P/C	TELLS: Guidelines for Remediation	1984
	Su/I/D	Manual for Interpreting Secondary School Reports	1984
	Su/I/D	Manual for Interpreting Intermediate School Reports	1984
	Su/I/D	Manual for Interpreting Elementary School Reports	1984
	M	PASCD Journal	Spring 1982
Rhode Island	St	Statewide Assessment Program: Basic Skills Testing Results and Life Skills Testing Results	1982-83 1983-84

BEST COPY AVAILABLE

STATE	CODE	REPORT TITLES	YEAR
South Carolina	St	Basic Skills Assessment Program. Cognitive Skills Assessment Battery: Preliminary Results	Fall 1984
	St	Basic Skills Assessment Program: Preliminary Report	Spring 1984
	St	Statewide Testing Program: Summary Report	1984
	C	Teaching and Testing Our Basic Skills Objectives (Reading): Grades 9-12	1983
	M	Measuring Educational Progress in the South: Student Achievement	1984
Tennessee	C	Proficiency Test Objectives. Their Domains with Sample Test Items	1983
	Tc	Statewide Assessment Program: Basic Skills Executive Summary and Basic Skills Technical Report	1982-83
Texas	St/Tc	Assessment of Basic Skills. Part I: Project Report. Part II: Technical Report	1982-83
	St	Assessment of Basic Skills: Statewide and Regional Results as Reported	1980-83
	C	Assessment of Basic Skills: Reading Objectives, Writing and Math Objectives, and Measurement Specifications (Grades 3,5,9)	1986
Utah	D/Tc	Educational Quality Indicators	1983
	M	An Analysis of Nation "Indicators of Risk"	1983
	St	Statewide Educational Assessment: General Report	1981
Virginia	St	Report on Public Education	1984
	St	Spring 1984 SRA Test Results	1984
	St	Minimum Competency Test Results	1982-83 Feb. 84
	Su	Statistical Data on Virginia's Public Schools	--
Washington	St/D	State General Report and District Level Summaries (reading, spelling, language arts, mathematics): Fourth Grade	Fall 1983
West Virginia	St	15th Report, State-County Testing Program	1982
	T	Student Questionnaire, Cognative Abilities Test, Level F, Form 3, Grade 9	1984

BEST COPY AVAILABLE

STATE	CODE	REPORT TITLES	YEAR
Wisconsin	St	Pupil Assessment Program Report	1977-83
Wyoming	M	Handbook for Establishing Minimum Competency Programs in Wyoming Schools	1982

KEY TO DOCUMENTS PROVIDED BY STATE TESTING PROGRAMS

Code	Category
C	Content Specification
D	District Summary
G	Group Right Response Reports
I	Item Report
IN	Instructional Guide
M	Misc.
P	Profile
SC	School Report
St	Statewide Report
Su	Support Materials (e.g. "interpreting results")
T	Test
Tc	Technical Report
TM	Test Administration Manual

APPENDIX 9

MASTER MATRICES
FOR MATH, READING, WRITING*

State:
Grade:
Source:
Year:
Test:

MATH

HIERARCHICAL PROCESS --->

SKILLS	RECALL	ROUTINE MANIPULATION	EXPLAIN TRANSLATE JUDGE	PROBLEM SOLVING	SKILL TOTALS
NUMBERS, NUMERATION	(10) o math facts o count o order o place value o symbols/word o # line o equiv. sets o equiv fract. o properties of #s o identity elements	(11) compute: o integers, o fractions o ratios o decimals o % o expanded notation o sequences o factors/mult o rounding o simple word problems o pos/neg #	(8) o computational estimation o know when to estimate o draw conclusion o ID assumption o select fact o sel. algorithm o sel. question o sel. problem modeled	(2) o est. in word problem o hard word probs: 2-step, %, interest, disct, finance charges	(31)
VARIABLES, RELATIONSHIPS	(3) o facts, def., o symb. of alg. <,>,= o laws of trig.	(4) o solve equalities & inequalities o read graphs o graph points/lines o complete function table	(2) o give equ. for given info o interpret formulas	(3) o solve probs w/ equations, trig o logic problems o graph problems	(12)
GEOMETRY... SIZE, SHAPE	(2) o def terms o recog. shape	(1) o find area, circumference, perimeter (simple)	(2) o translate words into symbol, fig o how fig looks from other	(4) o geom prob solvg o show 2 shapes congruent o apply theorems to solve probs o draw diagrams to solve problems	(9)

* Used to categorize & count test items & subskills. Each "o" indicates a subskill. This list is fairly comprehensive but does not contain every subskill tested.

BEST COPY AVAILABLE

State:
Grade:
Source:
Year:
Test:

MATH CONT.

SKILLS	RECALL	ROUTINE MANIPULATION	EXPLAIN TRANSLATE JUDGE	PROBLEM SOLVING	SKILL TOTALS
MEAS'MT, include maps, $, time, dist, weight, temp., etc.	(3) o def terms o equivalents o order	(3) o compute o conversions o reading instruments/measure	(3) o Identify most approp. unit to use o compare amts o est. sz of common things	(2) o word probs w/ measurement o estimate in word prob.	(11)
STATISTICS, PROBABILITY	(1) o def of terms*	(3) o compute mean, mode, median, range, etc. o organize data in table o compute probability	(1) o interpret* data	(1) o draw inferences* from data	(6)
TECHNOLOGY:* CALCULATIONS, COMPUTERS.	symb, terms flow charts Basic	read flow chart calculator computation	approp. time to use calc. & computer nonroutine computation	calc. application solve probs	
ATTITUDE*					
COGNITIVE TOTALS	(19)	(22)	(16)	(12)	(69)

* Did not occur on any tests.

BEST COPY AVAILABLE

State:
Grade Levels:
Source of Infor:
Year:
Test Used:

READING

HIERARCHY LEVEL

SKILLS	RECALL	LITERAL COMP.	INFERENTIAL EVALUATIVE COMPREHENSION	APPLIC'N	SKILL TOTALS
WORD ATTACK	(6) o phonetics o syllabication o affixes, roots o compound words o contractions o inflectional endings				(6)
VOCABULARY	(2) o meaning in isolation o signs	(3) o meaning in context o multi-meaning	(2) o analogy o nonsense in context		(6)
COMPREHENSION (note: content may be reg. para or "life skills," e.g., ads, etc.)		(7) o details o main idea o title o referents o sequence o cause/effect o follow directions	(20) o details/support o main idea/summary o title o irrel/miss'g info o missing words o sequence o cause/effect o conclusions o predictions o emot appeals o fact/opinion o A's purposes/attit. o A's methods o analyze character o figurative lang. o tone/emotion o contrast/compare o identify org. used o setting/plot/dialog o identify lit. type	(2) o select best X for given purpose o apply info to new situation	(29)

BEST COPY AVAILABLE

State:
Grade Level:
Source of Info:
Year:
Test used:

READING CONT.

HIERARCHICHAL LEVEL

SKILLS	RECALL ROUTINE MANIP.	EVALUATE	APPLICATION	SKILL TOTALS
STUDY SKILLS	(4) o Use info sources e.g., dic/guide words index/tab of c o Use card catalog o Use maps, charts o Alphabetize	(1) o identify which source to use		(5)
ATTITUDE				
COGNITIVE TOTAL	(8)	(23)	(2)	(46)

BEST COPY AVAILABLE

State:
Grade Levels:
Source of Info:
Year:
Test Used:

WRITING

SKILLS	RECALL	LITERAL COMP.	EXPLAIN INFER EVAL	APPLIC'N PROB SOLV	SKILL TOTALS
CONVENTIONS:		(7) o capitalize o punctuate o abbreviations o spelling o suffixes o plurals o contractions	(1) o when to ...	(see write sample)	(8)
GRAMMAR: (sentence structure)		(5) o parallel structure o complete sent. o compound, complex o subj/pred o parts of speech			(5)
WORD USAGE:		(6) o misplaced modifiers o language choices o subj/verb agreemt o transition words o dbl. negs o pronouns			(6)
ORGANIZATION:	(1) o Identify types of sentences	(9) o effective sent manip o sequence words o sequence sent. o seq paragraphs o select topic sent o sel important detail o sel info. to include o letter format o fill out forms	(2) o judge writing o edit re org'n		(12)
ATTITUDE:					
COGNITIVE TOTALS	(1)	(27)	(3)	---	(31)

WRITING SAMPLE:
Scoring: Holistic Primary Trait Analytic
Point syst:
Number/Type of writing sample: letter, theme story, other

Number of readers per sample:

BEST COPY AVAILABLE

APPENDIX 10

DECISION RULES

Test Analysis:

Math: for word problems--

NN-level 2:	simple word problems involving routine manip.
NN-level 4:	hard word problems, 2-step problems
G-level 4:	geometry problems including calculating area application, such as carpet or paint
M-level 4:	measurement problems other than the geometry ones above

Note: some reports did not make clear distinctions between subskills that were differentiated on the CSE matrices, such as the math example above. When tests were not available for analysis, it was necessary to rely on the categorization provided by the report.

Summary:

1. When tallying the number of items for summaries, if report says a groups of items includes some falling in 2 or more levels of the hierarchy of skills, divide equally for purposes of counting items. Be sure to count all subskills represented. E.g., report says there are 20 word problems including some that are 1-step (easy) and some 2-step (hard): count 10 items in level 2 of the hierarchy and 10 items in level 4. Also, count 1 subskill in level 2 and 1 in level 4.

2. When report did not mention number of items, only the number of subskills might be countable.

3. When report did not mention number of items or even which subskills of a skill area were tested, then only a check could be recorded for the subskill indicating that it was tested in some (unknown) fashion.

4. Note that there are many ways of dividing or grouping skills or objectives, and that the subskills used for classification purposes in this study are not necessarily "better" than some other scheme; they are just different and were useful for this study.

APPENDIX 11

RATING CATEGORIES OF "SOURCE QUALITY"

1. TEST (e.g. Montana, Illinois)

2. REPORT: straightforward, with clear, single skills and number of each type of items
   (e.g. Kansas, Louisiana, New Jersey, Missouri)

3. REPORT: reasonably good item specifications, however...
   a.) broad domains or clusters of skills that do not fit the subskills in the Content-by-Skill-Hierarchy Matrix, so cannot assign exact number of items per subskill even though the report is otherwise clear and may provide sample items.
   (e.g., California, Maryland)
   b. no info on number of items per subskill
   (e.g., Texas)

4. REPORT: list of "objectives" is too brief to be certain what items really measure; does provide number of items per objective.
   (e.g. Alabama)

5. REPORT: list of objectives, skills or domains very brief or vague; although may give a few sample items, report is not clear on what exactly is being measured. Does not provide number of items per objective.
   (e.g. Pennsylvania)

6a. REPORT: extremely vague or brief report mentioning only some of the skills tested, usually without information on the number of items on the test and without grade delineation.
(e.g. New Hampshire)

6b. INSTRUCTIONAL MATERIALS: vague as to what exactly is to be tested and no information on number of items per subskill.
(e.g. South Carolina)

6c. LETTER: mentions test exists but gives not specific information. May be a new program.
(e.g. Virginia, Mississippi)

NOTE: Some states provided different sources of information on different tests or content areas. In this case, more than one rating was given as appropriate.

NOTE: The above 6-point scale is ordinal only.

BEST COPY AVAILABLE

APPENDIX 12

Comments on Sources of Information and Quality of Information

Rating	State	Comment
5	ALABAMA - (Rpt.)	- Gives brief objectives and number of items (can't tell what items are really like)
1;5	ALASKA - 4th gr.test; 8th gr. - (Rpt.)	(skills mentioned but brief; no information on exact items)
	ARIZONA - (CAT)	
3a	ARKANSAS - (Rpt.)	- Report mentions appendix with list of objectives, but not sent to us. Report only lists major domains with # of subjects and items each - so isn't as helpful - can't tell how match our subskills, i.e., what's really measured.
3a	CALIFORNIA -(Rpt)	- (Broad domains = ours; can't assign #'s of items per subskill) - Good documentation otherwise; 12th grade = briefest; 8th grade most recent and best done re higher order skills.
	COLORADO	- no program
	CONNECTICUT	- no info
	DELAWARE	- CTBS
2-math 3-reading	FLORIDA - (Tech.)	- subskills easier to identify from report in math than in reading and writing.
5	GEORGIA - (Rpt.)	- very brief title of objs. so can't be sure what's measured or # of items.
	HAWAII - no info.	
2	IDAHO - (Rpt.)	- straightforward; # of items
1	ILLINOIS - (test)	- note: many items are same on 4th and 8th - and on 8th and 11th
4	INDIANA - (Rpt.)	- Very brief "obj" with item #'s - unsure what their "objs" really are. - Different items for grades 6 & 8 but areas are same and same # of items.

BEST COPY AVAILABLE

	IOWA - no prog.	
2	KANSAS - (Rpt. & list of objs)	- Straightforward with # of items per obj.
	KENTUCKY - CTBS-U	
2	LOUISIANA - (Legis. Rpt.) (Annual Rpt.)	- Straightforward with # of items per obj. - State assessment - Basic Skills
2	MAINE - (Summary Rpt.)	- Straightforward with # of items; extensive information on scoring of writing sample.
5	MARYLAND - (Rpt., Specs.)	- Specs = OK, but lump together several objs. under 1 domain - and don't give # of items.
	MASSACHUSETTS - no single statewide test; local choice	
1 for R&M 6c for Writing	MICHIGAN - (Test)	- Have writing objectives in Rpt. (?) - but no writing test -??
5	MINNESOTA Rept on MSRI Rept on MSEA-R Rept on MSEA-M Rept on Basic Math	- Confusing battery of tests No details on content for all tests - just brief "area" names which don't match ours well. Some item #s given.
2	MISSOURI - (Data Summary)	- Straightforward; gives #'s of items
1	MONTANA - (Test)	
6c	MISSISSIPPI (Rpt.)	- New program with no information on content other than RMW, grade levels.
	NEBRASKA - no prog.	
2	NEVADA - (Rpt.)	- Fairly straightforward "competency areas"; gives #'s of items.
6a	NEW HAMPSHIRE (Summary Rpt.)	- Vague: didn't give specific information on objs. or items and didn't differentiate grade levels by skill areas. Some areas and items mentioned in discussion of results (no list or tables, etc.).
2	NEW JERSEY ("Dir. of Specs & Items")	- Not real specs - but adequate for us. Gives # of items.

BEST COPY AVAILABLE

NEW MEXICO - CTBS-U

3a NEW YORK (Manuals) - Unique test of reading (infer missing words in prose passages). Math part of manual gives # of items in various content areas but uses different categories from ours - so can't assign # of items to our subskills.

5 NORTH CAROLINA (Rpt.) - Brief objs. only, no elaboration on content or # of items - a little hard to match to our categories/hierarchy.

NORTH DAKOTA - no prog.

OHIO - no prog.

OKLAHOMA - no prog.

3a OREGON (Summary Rpt) - Last few pages give # of items - but hard to make their categories match ours - theirs are large and vague, e.g., "inferential comp.", "evaluative comp."

5 PENNSYLVANIA

("EQA" Manual) - gives only brief name of item content - so tallies are tentative, especially on reading.

("TELLS" Booklet) - information only on "objs." - and brief' no # of items

RHODE ISLAND - ITBS

6b SOUTH CAROLINA ("Reading T&T 9-12") - Seems to be instructional manual, not specs or test manual; also, only covers 9-12 whereas test is done at 1-3, 6, 8, 11 - and also only reading, whereas test covers R,M, and W. Gives only areas of R tested, not # of items, and nothing on Math or writing [not very useful].

SOUTH DAKOTA - no prog.

5 TENNESSEE - (Rpt) - Gives obj. and some [sample?] items each, but doesn't specify # of items on test.

3b TEXAS - (Rpt.) - Gives reasonable, good specs and details on how specs and items written, but there are few objs covered; no information on # of items.

UTAH - CTBS-S

VERMONT - no prog.

BEST COPY AVAILABLE

6c VIRGINIA - (Rpt.) - Mentions there is minimum competency test in R&M at grade 10 - but gives no other information!

WASHINGTON - CAT

WEST VIRGINIA - no prog.

6a WISCONSIN - (Rpt) - Objs not listed. Only a few could be inferred from Rpt. # of items given only for whole test and "lit comp." subset.

WYOMING - no program.

APPENDIX 13

STQI PROJECT

Reading

Definition and Identification of Skills

CONTEXT x SKILL HIERARCHY MATRIX:

	RECALL	LITERAL COMP ROUTINE MANIP	INFER, JUDGE EXPLAIN	APPLIC'N
WORK ATTACK		no items	no items	no items
VOCABULARY				no items
COMPREHENSION	no items			
STUDY SKILLS	no items			no items

SAMPLE ITEMS

RECALL / WORD ATTACK:

1. PHONETICS

Look at the picture and the word under it. The word has missing letters. Choose the letters that are missing in the word.

* a. squ
b. spr
c. thr
d. shr

(picture of squirrel)

____irrel

2. SYLLABICATION

Look at the underlined word and select the response in which the word is correctly broken into syllables.

satisfaction

a. sat-is-fact-ion
b. satis-fac-tion
* c. sat-is-fac-tion
d. sa-tis-faction

3. AFFIXES & ROOTS

The root word in narrowing is:

* a. narrow
b. rowing
c. arrow
d. row

4. CONTRACTIONS

Which words mean the same as the underlined word?

You'll need an umbrella today.

  a. You all
  b. You would
  c. You still
* d. You will

5. INFLECTIONAL ENDINGS

Which underlined word shows that something happened in past?

When Eleanor arrives, you should show her the mural
a. b. c.

you painted.
* d.

RECALL / VOCABULARY:

1. MEANING IN ISOLATION

Which word means about the same as NOVICE?

  a. curator
  b. spendthrift
  c. weakling
* d. beginner

2. SIGNS

What does this sign mean?

  a. don't enter (stop sign)
* b. stop your car or bike
  c. stop talking
  d. no cars or bikes allowed

LIT. COMP. / VOCABULARY:

1. MEANING IN CONTEXT

Choose the word that means the same as the underlined word in the sentence.

Each morning Bernard has his customary breakfast of oatmeal, toast, and juice.

  a. fancy
  b. special
* c. usual
  d. strange

BEST COPY AVAILABLE

2. MULTIMEANING WORD

Choose the meaning of the underlined word as it is used in the sentence.

The snap has fallen off the collar of my shirt.

a. to make a sharp, crackling sound
b. a brief spell of cold weather
c. to snatch or grab suddenly
* d. a clasp on an article of clothing

INFER / VOCABULARY:

1. ANALOGY

Choose the word that best fits the blank.

SMALL is to LARGE as HIGH is to ________.

a. tall
b. tiny
* c. low
d. broad

2. NONSENSE IN CONTEXT

What is the best meaning for the underlined letters?

Sue mras kittens and puppies.

a. little
* b. likes
c. is
d. softly

ROUTINE MANIP (LITERAL COMPREHENSION / COMPREHENSION*:

1. DETAILS

(Given passage with explicitly stated detail...e.g.

A shock victim's skin is cold and may be moist to the touch. Pulse is fast and often too faint to be felt at the wrist. Breathing is rapid and shallow, and the victim feels weak and dizzy.

A person who is in shock is most likely to:

* a. feel dizzy
b. have a strong pulse
c. feel warm
d. take deep breaths

---

*Correct answers are not marked with an asterisk in items where reading passages have been omitted.

BEST COPY AVAILABLE

2. MAIN IDEA /

(Given passage with explicitly stated main idea...e.g.

At first glance, the prairie resembles little more than a barren and lonely expanse of grass, but in fact, the prairie is teeming with life. Among the most interesting inhabitants of the prairie are the harvester ants. Named for their habit of collecting seeds, these industrious insects are well suited to prairie life.
...(several more paragraphs about ants)

What is the main idea of this passage?

a. Harvester ants are well suited to life in the prairie.
b. Harvester ants' mounds are made of dirt.
c. Harvester ants hibernate during the winter.
d. A colony of harvester ants can collect a pint of seeds per day.

3. REFERENTS

(Given a passage, identify referent of a pronoun or word that functions like a pronoun.)

According to the story, who or what "sank slowly to the ground"?

a. the mule
b. the goat
c. the horse
d. the master

4. SEQUENCE

(Given passage, identify explicitly stated sequence of events.)

Which of the following happened last?

a. Jefferson became a musician
b. Jefferson wrote the Declaration of Independence
c. Jefferson was elected President
d. Jefferson designed his own home

5. CAUSE / EFFECT

(Given passage, identify explicitly stated cause or effect.)

Why did Linda stop in front of the house?

a. She saw a kitten
b. The children said the house was haunted
c. The house was old and big
d. She wanted to know what made the noise

6. FOLLOW DIRECTIONS — (E.g., given application form, identify correct way to fill it out according to written directions.)

On line 1, William should write the date on which he:

a. left his previous job
b. completes the application
c. began his first job
d. is available for work

INFER, EVALUATE / COMPREHENSION:

1. DETAILS, SUPPORT STATEMENTS — (Given passage,)

Which statement best supports James Lee's claim that the late bus would benefit students?

a. The school board should find a way to resume the services of the late bus
b. Extracurricular activities provide students with valuable learning experiences
c. Some students can get rides from their parents
d. Some working parents cannot take their children home from school

2. MAIN IDEA, SUMMARY, TITLE — (Given passage, infer best title, summary statement, title)

The main idea of these rules is that:

a. both adults and children enjoy the swimming pool
b. there is a snack bar at the swimming pool
c. safety is extremely important at the swimming pool
d. the swimming pool is open every day

3. MISSING / IRRELEVANT INFORMATION — (Given passage, infer missing information or identify important information to include or exclude)

Which of the following would be most important for the editors to include in this editorial?

a. The school has never given the band any money for its uniforms
b. Helmets and padding protect football players from injury
c. Members of the marching band perform indoor concerts too
d. The football team has longer practices than the marching band

BEST COPY AVAILABLE

4. MISSING WORDS — (Given reading passage with several words omitted, identify best word to fit in blank from context.) (Note: New York's entire reading test was like this)

5. SEQUENCE — (Given a passage, infers order of events or logic)

What indicates that Minnie was the first in her neighborhood to have a sewing machine?

a. The neighbor women all came to see it
b. She had to make everyone's clothes
c. Fred bought it
d. She didn't know how to operate it at first

6. CAUSE / EFFECT — (Given passage, infer cause or effect)

A major reason Paramount Studio moved to California was to -

a. allow the Army to use the Astoria plant
b. avoid the descruction of the studio by vandals
c. enable the Astoria plant to become a museum
d. be able to make movies less expensively

7. CONCLUSIONS — (Given passage, chart, etc., draw conclusions)

Based on the information in this chart, it may be concluded that:

a. cross-ventilation helps to warm a room
b. gas heat is more expensive than electric heat
c. fans use very little electricity
d. insulating walls conserves energy all year round

8. PREDICTIONS — (Given passage, predict probable outcome)

What probably happened next in this story?

a. The girl became angry and went home
b. Marina and the girl told each other their names
c. The girl made fun of Marina
d. Marina became embarrassed and stopped talking

9. FACT / OPINION

(Given passage or statement, distinguishes fact from opinion)

Which of the following is an example of an opinion?

a. "In 1860, a midwestern stagecoach company let people know about an exciting new plan."
b. "The mail must go through."
c. "The route cut directly across from Missouri to Sacramento."
d. "Each rider rode nonstop for about 100 miles."

10. PURPOSE, ATTITUDE

(Given passage, infer author's purpose or attitude)

The author's attitude toward the Pony Express riders can best be described as one of

a. confusion
b. amusement
c. worship
d. admiration

11. CHARACTER

(Given passage, identify character traits, identify motivations, draw conclusions about character's feelings)

The beasts and birds can best be described as

a. proud and closed-minded
b. understanding and wise
c. sleepy and lazy
d. thrifty, hard-working

12. FIGURATIVE LANGUAGE

(Given passage, identify meaning of metaphor, simile, idiom, or other image or figure of speech used)

The author's choice of words "sets up business" and "cleaning station" are used to show that

a. the wrasse's means of getting food is almost like a business service
b. wrasse fishing is big business
c. all fish set up stations
d. the wrasse enjoys cleaning itself in the water

BEST COPY AVAILABLE

13. TONE

(Given passage, recognize mood)

At the beginning of the story, the mood is one of

a. disappointment and sorrow
b. curiosity and excitement
c. fear and suspense
d. thankfulness and joy

14. COMPARE CONTRAST

(Given passage, infer similarities, differences)

Compared to American managers, Japanese baseball managers are -

a. better advisors
b. better paid
c. more knowledgeable
d. more powerful

15. ORGANIZATION

(Given passage, selection portion to complete outline or organizer based on organization of passage)

The following outline is based upon the last paragraph of the passage. Which topic below is needed to complete it?

I.
    A. Federalists
    B. Republicans

a. Competing parties
b. Jefferson's rivals
c. Election pay-offs
d. Strong governments

16. SETTING, PLOT DIALOGUE

(Given passage, identify and interpret time, place of story or event)

You can tell that this story took place

a. in a city park
b. at a zoo
c. in a forest
d. near a boot factory

17. LIT TYPE

(Given passage, recognize example of fiction, nonfiction, biography, autobiography, similes, metaphors, etc.)

The reading selection appear to be an example of

a. an autobiographical account
b. historical fiction
c. a biographical sketch
d. ancient mythology

APPLICATION / COMPREHENSION:

1. RELATE TO NEW SITUATION

(Given passage, relate ideas in it to situation not discussed)

Suppose your student council could not succeed in accomplishing any improvements for the student body because of many conflicts and divisions among the student members. Which of the following would be a way of applying Thomas Jefferson's beliefs to such a situation?

a. avoid the meetings so as not to waste time
b. try to unify the members to create an effective council
c. encourage the disagreements to create livelier debates
d. appoint one person to make all the decisions

ROUTINE MANIP. / STUDY SKILLS:

1. USE INFORMATION SOURCES

(includes dictionary entries and guide words, tables of contents, indexes, glossaries, encyclopedias, phone books, and other written information sources)

(Given a dictionary entry...)

Choose the definition that best fits how the underlined word is used in this sentence:

I can't trim your hair with these dull scissors.

a. v. 1
b. v. 2
c. n.
d. adj.

2. USE CARD CATALOG

(Given title card...)

Who is the author of Brother of the More Famous Jack?

a. Black Swan
b. Barbara Trapido
c. Victor Gollancz
d. Transworld Publishers

BEST COPY AVAILABLE

3. USE MAPS, CHARTS — (Given maps, charts, etc., to locate specific information or answer questions)

(Given chart of population of major U.S. cities...)

Which city had the least change in population between 1970 and 1980?

a. Philadelphia
b. Chicago
c. Houston
d. Los Angeles

4. ALPHABETIZE — Choose the word that comes first in alphabetical order.

a. solve
* b. sob
c. south
d. sort

EVALUATE, JUDGE / STUDY SKILLS:

1. SELECT BEST SOURCE — Where would you look to find a list of all the presidents of the United States?

* a. an encyclopedia
b. a newspaper
c. a dictionary
d. an atlas

ATTITUDE: I enjoy reading.

a. strongly agree
b. agree
c. not sure
d. disagree
e. strongly disagree

MATH SAMPLE ITEMS

Quality Indicators Project

Math

CONTEXT x SKILL HIERARCHY MATRIX:

	RECALL	ROUTINE MANIP.	EXPLAIN TRANSLATE JUDGE	PROBLEM SOLVING
NUMBERS, NUMERATION	1. order 2. number line 3. identity			
VARIABLES, RELATIONSHIPS				
GEOMETRY				
MEASUREMENT				
STATISTICS, PROBABILITY	(empty) i.e., no test items		(empty)	(empty)

SAMPLE ITEMS*

RECALL / NUMBERS:

1. ORDER What shows the correct relation of 7,9, & 16?

* a. 7 < 9 < 16 c. 16 < 9 < 7
b. 7 > 9 > 16 d. 16 > 9 < 7

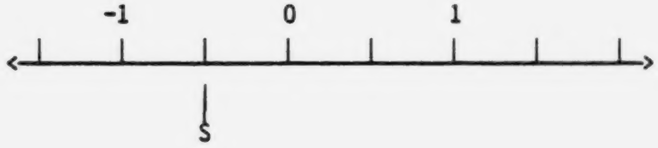
---

*Sample Items are presented for all cells in which test items occurred. Not every subskill in every cell is represented here, but the most frequent and characteristic ones are.

RECALL / NUMBERS (CONT.)

2. NUMBER LINE

What number is represented at point S on the number line?



* a. -1/2 c. 1/2
b. -1 1/2 d. 1

3. IDENTITY ELEM.

What value for "n" makes the sentence below true?

100 - n = 100

a. 0 * c. 1
b. 0.01 d. 100

4. MATH FACTS

4 x 2 =

* a. 8 c. 6
b. 9 d. 2

5. SYMBOLS/WORDS

Which number means "three hundred sixty-two"?

a. 352 * c. 362
b. 3620 d. 632

6. EQUIVALENT FRACTIONS

10/15 =

a. 3/5 c. 1/5
* b. 2/3 d. 3/2

ROUTINE MANIPULATION / NUMBERS:

1. COMPUTE: WHOLE NUMBERS

79 + 34 =

a. 112 c. 103
b. 45 * d. 113

2. FACTORS, MULTIPLES

Which shows the prime factorization of 12?

a. 3 x 4
b. 1 x 12
* c. 3×2^2
d. $2 \times 3 \times 2^2$

3. NUMBER SEQUENCES

Which number is missing? 1011, 1022, ____, 1044

a. 1043
* b. 1033
c. 1023
d. 1020

4. SIMPLE WORD PROBLEMS

A basketball team has won its first 3 games. It must play 12 games in all. What percent of the total games has the team played?

* a. 25%
b. 3%
c. 33%
d. 75%

5. ROUNDING

Round 0.4088 to the nearest hundredth.

a. 0.40
b. 0.408
c. 0.409
* d. 0.41

6. CONVERT FRACTIONS, DECIMALS, %

3/4 =

* a. .75
b. .34
c. 3.4
d. 75.0

EXPLAIN, JUDGE / NUMBERS

1. FORMULATE PROBLEM

JoAnn works 4 hrs a day for 4 days a week. She earns $4.25 an hour. She wants to earn enough money to buy a refrigerator for $585.

Which problem cannot be solved with the information given above?

a. How much money does JoAnn earn each week?
b. How many days must JoAnn work to buy the refrigerator?
c. How much more money would JoAnn earn each week if she is paid $5.00 an hour?
* d. What is the capacity of the refrigerator that JoAnn will buy?

2. IDENTIFY FACTS

Joe bought a shirt that regularly sells for $24 on sale for $18. What percent off the regular price was the sale price?

What facts are given?

- a. sale price and discount rate
- * b. sale price and regular price
- c. regular price and discount rate
- d. regular price, selling price, and discount rate

3. IDENTIFY ALGORITHM

A packet of gelatin weighs 20 grams. What is the weight of 10 packets of gelatin?

Which of the following problems can be solved using the same operations as the problem above?

- a. Juanita runs 10 miles in 90 min. How long does it take her to run each mile?
- b. A felt pen costs 49¢ and a ballpoint costs 99¢. How much does a felt pen and a ballpoint cost?
- c. It takes 4 ounces of juice to fill a glass. How many glasses can be filled from a half-gallon bottle of juice?
- * d. A pencil costs 10¢. What is the cost of 4 dozen pencils?

4. EVALUATE CONCLUSIONS, ASSUMPTIONS

Magdelena got 80% correct on a math test and 85% correct on a science test. Ralph said that Magdelena got more right answers in the science test than in the math test.

Which of these conclusions about Ralph's statement is correct?

- a. Ralph's statement is true under all conditions.
- b. Ralph's statement cannot be true under any condition.
- * c. Ralph's statement is true if the tests each have the same number of questions.
- d. Ralph's statement cannot be true if the tests each have the same number of questions.

5. COMPUTATIONAL ESTIMATION

Estimate the product: 89.61 x 10.42

a. 9000 b. 1200 * c. 900 d. 100

BEST COPY AVAILABLE

PROBLEM SOLVING / NUMBERS:

1. ESTIMATE IN WORD PROBLEM

The payroll of a grocery store for its 23 clerks is $395,421. What is the best estimate of the answer?

What is the best estimate of the answer?

* a. $20,000
b. $17,192.22
c. $20.00
d. $1300

2. HARD WORD PROBLEMS OR 2-STEP PROBLEMS

With 5 games to play, Steve had 187 hits. In his next four games, he got 1,4,2, and 3 hits. How many hits must he get in his last game to have a 200-hit season?

a. 2
* b. 3
c. 10
d. 13

RECALL / VARIABLES:

1. SYMBOLS

Choose the symbol that makes the number sentence true:

$$3 + 4 \ \square \ 8$$

a. -
b. >
* c. <
d. =

ROUTINE MANIP. / VARIABLES:

1. EQUATIONS, INEQUALITIES

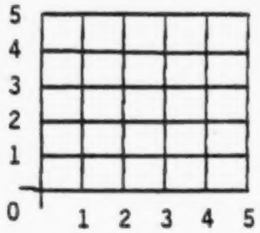
If x is replaced by 3, then the value of $x^2 - 1$ is

a. 2
b. 5
* c. 8
d. 11

2. GRAPH POINTS

The point F is named by:

* a. (2,3)
b. (3,2)
c. (3,3)
4. (2,4)



BEST COPY AVAILABLE

3. READ GRAPHS (given a line graph . . .)

In what year did the Tigers win 15 games?

a. 1980 b. '81 c. '82 * d. '84

EXPLAIN, JUDGE / VARIABLES:

1. ORGANIZE DATA IN GRAPH, TABLE

Put these test scores into a frequency table:

Sue scored 5, John--6, Tony--9, Sarah--10, Tom--6, Brad--7, Jenny--9, Kate--6.

2. EQUATION FOR INFORAMTION

John and Tom have 10 books between them. Tom has 2 more books than John does.

Which pair of equations describe this information?

* a. $J + T = 10$, $J + 2 = T$
b. $J = T = 10$, $J + 2 = T$
c. $J + T = 10$, $J = 2 + T$
d. $J - T = 10$, $J - 2 = T$

PROBLEM SOLVING / VARIABLES:

1. GRAPH PROBLEM

(given a line graph of number of arrests by years . . .)

During which 2 years were the same number of people arrested for drunken driving?

* a. 1975 & 1976 c. '77 & '78
b. '78 & '80 d. '77 & '79

2. FORMULA

Find the volume of the pyramid with a rectangular base using the formula $V = Bh/3$ (B = area of base, h = height)

a. 30 cubic inches
b. 32 cubic inches
c. 90 cubic inches
d. 96 cubic inches

RECALL / GEOMETRY:

1. DEFINE TERMS

Which figure shows a ray?

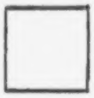
a. c.
b. * d.

2. CONCEPTS

This figure is a square:

What is the measure of angle I?

a. 30
b. 45
c. 60
* d. 90

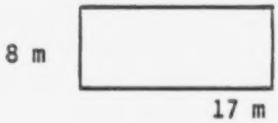


ROUTINE MANIPULATION / GEOMETRY:

1. AREA, CIRCUM, PERIM, VOL. SIMPLE

Find the area of the rectangle below:

a. 25 m
b. 42 m
c. 68 m
* d. 136 m



2. CORRESPONDING SIDES, ANGLES

Given that the figures below are similar, the measure of F is the same as the measure of

a. H
* b. M
c. N
d. O

3. SHAPES

The figure below shows the part of a figure on one side of a line of symmetry, m. Which answer choice shows the complete figure?

a. b. * c. d.

JUDGE, EXPLAIN / GEOMETRY:

1. SHAPES FROM OTHER VIEW

Figure F below shows a block with one corner cut off and shaded. Which answer shows a figure of how this block would look when viewed from directly above it?

a. c.

* b. d.

Fig. F

2. ESTIMATE

Estimate the size of the angle below. It appears to be between:

* a. 0 and 45
  b. 45 and 60
  c. 90 and 135
  d. 135 and 180

PROBLEM SOLVING / GEOMETRY:

1. APPLY THEOREMS

Which of the following statements is true about a square and a triangle both inscribed in the same circle?

  a. The area of the square is greater than the area of the triangle.
  b. The square and the triangle have the same perimeter.
* c. The arc of the triangle is greater than the arc of the square.
  d. The perimeter of the triangle is greater than the perimeter of the square.

2. WORD PROBLEMS

Robert must choose one of 4 solid chocolate candies to buy. Which one of the following shapes will give him the MOST chocolate for his money?

* a. Cube one inch on a side.
  b. Sphere one inch diameter
  c. Cylinder one inch in height and one inch in diameter
  d. Pyramid one inch in height with a one inch square base
  e. Cone one inch in height and one inch in diameter

RECALL / MEASUREMENT:

1. EQUIVALENTS

How many inches equal one yard?

  a. 30
  b. 35
* c. 36
  d. 39

2. ORDER

Which month comes next after April?

* a. May   b. March   c. June   d. February

ROUTINE MANIPULATION / MEASUREMENT:

1. READ INSTRUMENTS

What time is it?

a. 3:00
* b. 3:30
c. 4:00
d. 4:30

2. CONVERSIONS

One meter equals

* a. 39.14 in.
b. 36 in.
c. 3 yards
d. 41 in.

EXPLAIN / MEASUREMENT:

1. IDENTIFY BEST UNIT TO USE

Which unit is best for measuring the distance between two cities?

* a. kilometer
b. centimeter
c. liter
d. kilogram

2. ESTIMATE

Which object would be about 4 meters long?

a. bicycle
* b. automobile
c. shoe
d. baseball bat

PROBLEM SOLVING / MEASUREMENT:

1. WORD PROBLEMS OBJECTS

A map of a state is to be drawn so that one-fourth inch represents five miles. If the real distance between two points in the state is 20 miles, how many inches apart should these two points be on the map?

a. 1/2 inch
b. 3/4 inch
* c. 1 inch
d. 1 1/4 inch

2. ESTIMATE MEASUREMENTS IN WORD PROB

(Given a map . . .)
Using Routes 21 and 222, what is the approximate distance from Crestline to Pleasantburg?

a. 12 mi.
* b. 20 mi.
c. 30 mi.
d. 35 mi.

BEST COPY AVAILABLE

RECALL / STATISTICS:

NONE

ROUTINE MANIP. / STATISTICS:

1. COMPUTE MEAN, MEDIAN, MODE

From Monday through Thursday, Roman's News Stand sold 17, 36, 41, and 30 copies of the Town News. What was the average number of papers sold per day?

* a. 31 c. 114
b. 33 d. 124

2. COMPUTE PROBABILITIES

The sectors of the spinners are colored red (R), blue (B), and green (G). What is the probability that the spinner will stop on the blue if you spin it one time?

a. 1/2
b. 2/3
c. 1/4
* d. 1.3

EXPLAIN / STATISTICS:

NONE

PROBLEM SOLVING / STATISTICS:

(DRAW INFERENCES--NONE)

STQI PROJECT

Writing

Definition and Identification of Skills

CONTENT x SKILL HIERARCHY MATRIX

	RECALL	ROUTINE MANIP	INFER,EVAL	APPLLIC'N
CONVENTIONS	no items		no items	(see writing samples)
GRAMMAR			no items	"
WORD USAGE	no items		no items	"
ORGANIZATION	no items			"

SAMPLE ITEMS

ROUTINE MANIP./ CONVENTIONS:

1. CAPITALIZATION — Mark the answer that completes the sentence correctly. The longest river in the United States is the ____________.
 a. Mississippi river
 b. mississippi river
 *c. Mississippi River
 d. mississippi River

2. PUNCTUATION — Mark the answer that completes the sentence correctly. Our high school band includes ______ trumpets, and drums.
 a. clarinets
 b. clarinets;
 *c. clarinets,
 d. clarinets.

3. ABBREVIATIONS — The abbreviation for "street" is:
 *a. st.
 b. st
 c. stt
 d. s.

4. SPELLING — Choose the correct spelling of "9"
 a. nin b. nien *c. nine d. nein

5. SUFFIXES, AFFIXES — Choose the letter or letters needed to spell the word correctly.
 We will go swim_____ every day.
 a. ing *b. ming c. eing d. in

6. PLURALS — Choose the word which completes the sentence correctly.
 My two front ______ are missing
 a. tooths *b. teeth c. teeths d. tooth

7. CONTRACTIONS

Choose the word which completes the sentence correctly. I ________ seen her all day.
a. hav'ent b. hav'nt *c haven't d. havent

RECALL/GRAMMAR:

1. SENTENCE TYPES

Choose the interrogative sentence:
*a. What should we do about it?
b. Let's go to the store in an hour.
c. What a sight that must have been!
d. Marina checked out the book I wanted.

ROUTINE MANIP./GRAMMAR:

1. COMPLETE SENTENCES

Choose the one which will form one or more complete sentences.
We go camping to get away from ____.
a. crowds. To enjoy the peace and quiet.
b. crowds, we enjoy the peace and quiet.
*c. crowds. We enjoy the peace and quiet.
d. crowds. Enjoying the peace and quiet.

2. SUBJECT, PREDICATE

Choose the one which will form one or more complete sentences.
The school carnival ____________________.
a. next week
b. games and prizes
c. lots of fun
*d. is coming

3. COMPOUND OR COMPLEX SENTENCES

Choose the one below which combines the numbered sentences in the best way.

1. Ladybugs are beatles
2. Ladybugs are small
3. They feed on insects

* a. Ladybugs are small beetles that feed on insects.
b. Ladybugs are beetles, and they are small, and they feed on insects.
c. Ladybugs feed on insects, and they are beetles, and they are small.

4. MISPLACED MODIFERS

Which of the following revisions, if any, corrects the grammar in this sentence:
You can call your mother in London and tell her all about George's taking you out to dinner for just sixty cents.
*a. Move "for just sixty cents" to the beginning.
b. Change "George's" to "George"
c. Change "can call" to "could call"
d. Move "in London" to the end.

5. PARALLELISM

Mark the letter for the location of the error in this sentence:

Students in our $\underset{\text{a.}}{\underline{\text{French}}}$ class $\underset{\text{b.}}{\underline{\text{like}}}$ reading $\underset{\text{c.}}{\underline{\text{better}}}$ than $\underset{\text{*d.}}{\underline{\text{to work}}}$.

ROUTINE MANIP./WORD USAGE:

1.	LANGUAGE CHOICES (specificity, senses, tone)	Select the one which suggests an unfriendly attitude from Mr. Houser. Mr. Houser ________ that we pay the bill. a. asked *b. demanded c. requested
2.	SUBJECT-VERB AGREEMENT	Mark the letter for the location of the error. Because Tyrone is $\underset{\text{a.}}{\underline{\text{really}}}$ afraid of snakes, $\underset{\text{b}}{\underline{\text{he}}}$ $\underset{\text{*c}}{\underline{\text{don't}}}$ want to go hiking with $\underset{\text{d}}{\underline{\text{us}}}$.
3.	TRANSITION WORDS	Choose the word that best completes each sentence. You may use the same word more than once. To be a skillful debater, you must be able to argue both sides of an issue. (1) ______ study the side that you will defend. (2)_______ test your position with arguments from the opposing side. (3) _____ this may become a tedious task, it is usually the most prepared debater who wins. a. Then b. First c. Although d. Otherwise (2) (1) (3)
4.	DOUBLE NEGATIONS	Choose the one that completes the sentence correctly. He didn't buy ______ popcorn. a. no *b. any c. none
5.	PRONOUNS	Mark the letter for the location of the error. He spoke $\underset{\text{a}}{\underline{\text{bluntly}}}$ and $\underset{\text{b}}{\underline{\text{angrily}}}$ to $\underset{\text{*c}}{\underline{\text{we}}}$ spectators.
6.	VERB FORMS	Choose the one that completes the sentence correctly. Every day I walk to work, but Bob ____ . a. run *b. runs c. runned d. ran

ROUTINE MANIP./ORGANIZATION:

1.	SENTENCE MANIPULATION	Mark the sentence below which expresses the thought most effectively and econmically. a. He spoke to me in a very warm manner when we met each other Tuesday. b. When we met Tuesday, I was spoken to in a very warm manner by him. c. His manner was very warm when meeting and speaking to me Tuesday. d. Tuesday he greeted me warmly.
2.	SEQUENCE SENTENCES	Choose the best order to arrange sentences into a logical paragraph. 1. At the first traffic light, you'll see a red brick house on the corner.

BEST COPY AVAILABLE

2. To get to my house, turn right after you leave the school and walk straight for three blocks.
3. Walk down that street until you see a house with a blue porch --- that's my house.
4. Turn left there.

*a. 2-1-4-3 b. 2-3-1-4- c. 3-4-1-2 d. 1-2-3-4

3. SELECT TOPIC SENTENCE

Choose the sentence which is the best topic sentence (main idea) for the paragraph.

____
______________. You should try to stay away from trees and telephone wires...(paragraph continues)

a. It is so much fun to make a Kite
*b. When you're flying a Kite, there are several things you should keep in mind.
c. It is so much fun to fly a kite.
d. When you're buying a kite, you should remember to take enough money with you.

4. SELECT IMPORTANT DETAIL

Choose the best supporting detail for the main idea expressed by the sentence:
My youngest brother was frightened on his first day of school.

a. My father was afraid of school when he was younger.
b. He already knew the alphabet.
*c. He cried and clung to my father's hand.
d. The teacher was friendly and encouraging.

5. SELECT INFO TO INCLUDE

The following outline was used in writing the paragraph below it. Choose the sentence needed to complete the paragraph according to the outline.

I. Athletes don't get fat
   A. Example _____ tennis players
   B. Other examples ____ gymnasts and wrestlers
   C. Conclusion _____ strict diets

Most successful athletes don't allow themselves to become fat, because extra weight slows them down. __________ __________. If they are 10 pounds overweight, they may be slowed down...(para. continues)

a. There are many sports which I enjoy watching.
*b. Tennis players, for example, have to move with lightning speed.
c. You can play tennis at any age.
d. Staying on a diet is difficult.

6. LETTER FORMAT

Mark the letter for the location of the error.
(Given letter with underlined elements...)
*a. (lack of complimentary close)

APPLICATION/ORGANIZATION:

1. EDIT ORG'N

You are to make decisions about what should be revised to improve the selection below. The underlined sentences are the ones about which there are questions. Use they KEY below to make judgments about each of the sentences.

What is your best decision about the underlined, numbered sentences?

KEY:

A. KEEP. It is all right where it is.
B. TAKE OUT. It doesn't fit anywhere.
C. CHANGE. It is not clear at all and should be said in another way.
D. MOVE. It should be at another place.

(Given paragraph with underlined sentences...)

2. JUDGE WRITING

Read the student letter, and answer the question below.

Dear Mr. Vega,

I think the tidal pools would be a fun place to go for the fifth graders. It would be very interesting and fun. Please consider this request carefully.

Yours truly,
Pat Jones

Suppose your friend just wrote this letter. What advice would help her make it more convincing to the principal?

a. Indent "Dear Mr. Vega."
b. Add Mr. Vega's address in the upper right-hand corner of the letter.
c. Mention the dangers of going to the tidal pools.
*d. Add examples of what could be learned by going.

ATTITUDE:

1. Good writing is important to me because it helps me to get good grades.
a. strongly agree
b. agree
c. neither agree nor disagree
d. disagree
e. strongly disagree

2. Good writing will help me learn to express myself.
a. very unlikely
b. unlikely
c. neither likely nor unlikely
d. likely
e. very likely

BEST COPY AVAILABLE

WRITING SAMPLE

This part of the writing test consists of one writing exercise in which you will be expected to show how well you can write. For the exercise, you will write an essay on the stated topic.

You will have 30 minutes to complete your essay. You may wish to take the first few minutes to think about how you will organize what you have to say before you begin to write. If you wish to make an outline or any notes, use the space for notes provided on the back of this sheet. This space is meant to help you plan your essay, but your notes will not be scored. All that will be scored is the essay you write on the 2 lined pages provided.

Do your best to write a clear, well organized essay. You may not use a dictionary or any other reference materials during the test. If you finish your essay before time is called, read what you have written and make any changes that you feel will improve your writing.

TOPIC: Think of something important that happened in your life. It may have been happy or sad, painful or enjoyable. Write an essay in which you tell what happened and why it was important to you.

APPENDIX 14

KEY TO SUMMARY SHEETS
MATH, READING, AND WRITING

ENTRIES IN TABLE

In some cases, both the number of items and the number of subskills are known, in which case both appear in the table.

Numbers on the left of the slash indicate the number of items on the test that fall in that row or column of the matrix.

Numbers on the right side of the slash indicate the number of different subskills from that row or column of the Master Matrix that are tested.

When the number of items is unknown, only the number of subskills (the number on the right of the slash) appears in the table.

When neither the number of items nor the number of subskills is known, a "?" appears in the table if the state's materials mentioned that at least one subskill in that row or column is on their test.

MATH HEADINGS

Skill areas:
N = Numbers & numeration (symbols, properties, computation, word problems)
V = Variables & relationships (algebra, trig, graphing)
G = Geometry (terms, shapes, formulas, theorems, word problems)
M = Measurement (metric & US Customary units: terms, coversion, word problems)
S = Statistics & Probability (computation, problems)

Hierarchy level:
1 = Recall (facts, definitions, symbols, concepts)
2 = Routine Manipulation (basic computation, manipulation)
3 = Explain, Translate, Judge (evaluate, attention to process)
4 = Problems Solving (apply concepts, operations & facts, word problems)

READING HEADINGS

The headings in Reading and Writing combine content and skill hierarchy since a number of the cells in the full matrix were not tested by any state, according to their materials.

WA = Word Attack (First or Recall level; includes phonetics, affixes, syllabication, etc.)
VOC = Vocabulary (Spread across the Recall level second or literal comp level, and third or Infer level of the hierarchy. Preponderance of items were at 2nd level.)

LC = Literal Comprehension (Second or Routine Manipulation/Literal level)
IC = Inferential, Evaluative Comprehension (Third level, except for a single subskill involving application of reading to new context...4th level)

SS = Study Skills (Primarily at the Second level, using information sources; one subskill--judging which sources is appropriate--is at the Third level)

AT - Attitude toward reading (no level specified)

WRITING HEADINGS

CO = Writing conventions (e.g. spelling, capitalization, punctuation, at the Second level)

GR = Grammar (sentence structure, parts of speech, etc., at the first and second levels)

WU = Word Usage (choice of correct or most appropriate words, at the second level)

OR = Organization (effective sentence manipulation, organization of words, sentences and paragraphs, all at the second level except 2 subskills involving editing or judging organization)

AT = Attitude towards writing (no level specified)

SM = Writing Sample (e.g. letter, theme, at the 4th or application level, cutting across the above content.)

BEST COPY AVAILABLE

APPENDIX 15

READING

LIST OF STATES FOR STQI PROJECT

				1 - 3							4 - 6						
		STATE		WA	VOC	LC	IC	SS	AT		WA	VOC	LC	IC	SS	AT	Sub-skill
4	26	ALABAMA	CRT	32/8	12/6	16/4	16/4	16/4	—	CAT	12/3	20/5	16/4	12/3	12/3	—	18
			CAT	36/5	15/1	6/2	14/4	—	—		—	30/2	8/2	32/8	35/3	—	15
1		ALASKA									12/4	3/1	7/6	11/7	12/3	—	21
	12	ARIZONA	CAT	36/5	15/1	6/2	14/4	—	—	CAT	—	30/2	8/2	32/8	35/3	—	15
3a	27	ARKANSAS	CRT	36/9	—	36/9		36/9	—	SRA	24/6	—	36/9	36/9	56/14	—	29
			SRA								—	30/1	9/2	17/7	6/1	--	11
3	14	CALIFORNIA		60/3	30/2	73/3	77/4	30/2	—		16/1	54/2	62/3	78/16	30/4	—	26
		COLORADO															
		CONNECTICUT	OLD							CAEP	--	11/3	12/4	34/11	30/5	13	23
			NEW							4th	--	—	12/3	24/11	11/4	—	18
										6th	—	--	/3	/16	/4	—	23
	8	DELAWARE	CTBS/U	30/1	25/3	21/3	4/1	—	--		5/1	40/2	14/4	29/8	20/4	—	19
3	8	FLORIDA		5/1	15/1	13/3	10/2	5/1	—		9/1	10/1	24/4	5/1	--	--	7
5	6	GEORGIA		—	—	?2	?4	—	--		—	?	?2	?4	--	—	6
		HAWAII	SAT	72/2	38/1	30/4	30/7	10/2	—		60/1	36/1	30/4	30/7	20/3	—	16
			CRT	—— no info. ——													
		IDAHO															
1		ILLINOIS									—	8/1	9/2	4/4	--	--	7
4	14	INDIANA		—	15/3	25/5	30/6	--	--		—	15/3	25/5	35/7	—	--	15
		IOWA															
2	11	KANSAS		18/2	6/2	9/3	9/3	3/1	—	4th	15/5	9/3	6/2	12/4	18/6	—	20
										6th	12/3	9/2	9/3	12/4	18/6	--	18
	8	KENTUCKY		30/1	25/3	21/3	4/1	--	--		5/1	40/2	14/4	29/8	20/4	--	19
	10		2nd	8/1	16/2	20/5	—	12/2	—		20/4	20/1	4/1	12/3	16/4	—	13
2	14	LOUISIANA	3rd	36/5	4/1	20/5	—	12/3									
2		MAINE									—	--	15/3	13/4	12/4	—	11
4	12	MARYLAND	CAT	36/5	15/1	6/2	14/4	--	--	CAT	—	30/2	8/2	32/8	35/3	—	15
		MASSACHUSETTS				Local							Local				
1		MICHIGAN									6/1	15/3	18/4	27/8	9/4	15	21

KEY:
WA = Word Attack
VOC = Vocabulary
LC = Literal Comprehension
IC = Inferential Comprehension
SS = Study Skills
AT = Attitude
# of items/# of subskills
? = unknown # of items and subskills

BEST COPY AVAILABLE

READING

				1 - 3						4 - 6						
		STATE		WA	VOC	LC	IC	SS	AT	WA	VOC	LC	IC	SS	AT	Sub-skill
6c		MISSISSIPPI		NEW						NEW						
5		MINNESOTA								64/3	41/4	37/3	9(?)	20/3	—	(?)
2		MISSOURI								--	--	2/2	9/3	2/2	--	7
1		MONTANA								36/6	18/2	8/4	26/4	33/4	15	21
		NEBRASKA		No Program												
2	16	NEVADA	SAT	72/2	38/1	30/4	30/7	10/2	—	601	36/1	30/4	30/7	12/3	--	16
6a		NEW HAMPSHIRE								?	?	?	?	?3	--	(?)
2		NEW JERSEY		Local Choice						Local Choice						
	8	NEW MEXICO	CTBS/U	30/1	25/3	21/3	4/1	—	--	5/1	40/2	14/4	29/8	20/4	--	19
3a	1	NEW YORK	"DRP" (infer missing word)	—	—	--	56/1	—	--	--	--	--	77/1	--	--	/1
	12	NORTH CAROLINA	CAT	36/5	15/1	6/2	14/4	—	—	0	30/2	8/2	32/8	35/3	--	15
		NORTH DAKOTA		No program												
		OHIO		No program												
		OKLAHOMA		No program												
3a		OREGON								19/5	9/2	16/4	4/3	12/5	--	/19
5	12	PENNSYLVANIA		—	?2	?5	?4	?1	—	--	?2	?2	?6	?4	—	/14
		RHODE ISLAND										ITBS	(4,6)			
6	14	SOUTH CAROLINA	CRT	?	/2	/2	/8	/2	--	?	?	/2	/8	/2	—	/12
			CTBS/U							5/1	40/2	14/4	29/8	20/4	--	19
		SOUTH DAKOTA		No program												
		TENNESSEE														
36	9	TEXAS		/2	/2	/3	/1	/1	—	--	/1	/2	/3	/2	—	9
		UTAH	CTBS/S							—	40/1	12/3	33/9	20/3	--	16
		VERMONT		No program												
		VIRGINIA	SRA							--	30/1	9/2	17/7	6/1	--	11
		WASHINGTON	CAT							--	30/2	8/2	32/8	35/3	—	15
		WEST VIRGINIA		30/1	25/3	21/3	4/1	—	—	5/1	40/2	14/4	29/8	20/4	—	19
6a		WISCONSIN	CTBS/U							--	—	/7	/5	/2	--	/14
			CTBS/U							5/1	40/2	14/4	29/8	20/4	—	19
		WYOMING		No program												

BEST COPY AVAILABLE

READING

LIST OF STATES FOR QUALITY INDICATORS PROJECT

						7 - 9							10 - 12				
		STATE		WA	VOC	LC	IC	SS	AT		WA	VOC	LC	IC	SS	AT	Sub-skill
4	19	ALABAMA	CRT	8/2	12/3	12/3	16/5	24/6	—		8/2	13/3	11/2	18/4	30/6	—	17
	17		CAT	—	30/2	8/2	32/11	15/2	—				No Information				
5		ALASKA		?	?	?	?	?	—								
		ARIZONA				CAT							CAT				
3a	23	ARKANSAS		--	16/4	24/6	16/4	36/9	--								
3a	23	CALIFORNIA		15/1	68/2	48/3	235/15	36/2	—		--	3/1	47/4	50/5	13/2	—	12
		COLORADO															
		CONNECTICUT	CAEP	1/1	26/2	27/5	285/16	54/9	21		1/1	26/2	27/5	285/16	54/9	21	33
			Prof.	--	3/1	—	3/1	8/4	—								
			Master	—	--	/3	/16	/4									
	20	DELAWARE	CTBS/U	5/1	40/2	2/1	43/13	20/3	—				CTBT				
3	?	FLORIDA[I]*		15/1	10/1	18/3	9/2	14/?	—	[1]*	5/1	5/1	5/5	15/5	--	--	--
										[2]*	--	--	20/4	15/3	20/3	—	14
5	8	GEORGIA		—	?	?2	?6	—	—		--	--	?2	?10	?3	—	15
		HAWAII			SAT & DAT							STAS					
2	14	IDAHO		—	3/1	25/6	14/4	22/3	—	CRT	—	/1	/1	/1	/6	—	9
1	11	ILLINOIS		—	10/2	10/2	10/7	—	--		--	13/2	6/2	7/3	--	--	7
4	15	INDIANA		—	15/3	25/5	35/7	—	--								
		IOWA															
2	15	KANSAS		3/1	6/2	18/4	15/5	15/3	—		--	3/1	30/2	21/7	6/2	--	12
	20	KENTUCKY		5/1	40/2	2/1	43/13	20/3	—				CTBS/U				
2	16	LOUISIANA		24/5	8/1	20/5	15/3	8/2	--		8/2	12/3	20/5	24/6	8/2	—	18
2	11	MAINE		—	—	13/3	15/4	12/4	—		—	--	10/3	16/3	12/4	—	10
5	?	MARYLAND											CAT				
		MASSACHUSETTS															
1	20	MICHIGAN		3/1	15/3	16/5	24/7	9/3	15		3/1	15/3	21/5	26/8	7/4	15	24

---

*Florida [1] = State Student Assessment Test - Part I
[2] = State Student Assessment Test - Part II

BEST COPY AVAILABLE

READING

LIST OF STATES FOR QUALITY INDICATORS PROJECT

			7 - 9						10 - 12						
		STATE	WA	VOC	LC	IC	SS	AT	WA	VOC	LC	IC	SS	AT	Sub-skill
	?	[1]	54/2	74/4	40/2	23/?	20/?	—							
5	15	MINNESOTA*[2]	30/3	30/3	33/2	15/3	18/3	—	54/2	71/4	48/2+	24/?	26/?	—	( )
															[It = 223]
6c		MISSISSIPPI			NEW						NEW				
2		MISSOURI							--	—	—	12/6	--	--	/6
1		MONTANA							--	25/2	1/1	20/6	30/4	15	14
		NEBRASKA		No Program											
2	10	NEVADA	—	—	24/4	12/2	24/4	—	Same (9-12 High School Prof. exam)						10
6a		NEW HAMPSHIRE	?( )	?	?	?	?	--	?()	?	?	?	?	--	( )
2	20	NEW JERSEY in	—	12/1	13/4	43/11	22/4	—	Same (9-12 exam)						20
	25	out	15/5	20/1	21/4	34/10	20/5	--							
	20	NEW MEXICO CTBS/U	5/1	40/1	2/1	43/13	20/3	—			CTBS				
3a	1	NEW YORK	—	—	—	77/1	—	--	--	--	—	77/1	--	--	/1
5	17	NORTH CAROLINA CAT	—	30/2	8/2	32/11	15/2	—	--	/1	/3	/4	/3	—	11
		NORTH DAKOTA		No program											
		OHIO		No program											
		OKLAHOMA		No program											
3a	18	OREGON	6/2	13/2	21/5	6/5	13/4	—	--	15/2	10/2	15(?)	20/4		( )
5	16	PENNSYLVANIA	—	?3	?2	?7	?4	--	--	3/1	7/2	34/7	--	--	10**
		RHODE ISLAND			ITBS										
6	12	SOUTH CAROLINA CRT	?	?	?2	?8	?2	--	?	?	?2	?9	?2	—	/13
	20	CTBS/U	5/1	40/2	2/1	43/13	20/3	--							
		SOUTH DAKOTA		No program											
5	15	TENNESSEE	?2	?2	?5	?4	?2	—	Same (9-12 exam)						/15
3b	11	TEXAS	--	?1	?2	?6	?2	--							
		UTAH CTBS/S							--	40/1	3/1	43/9	20/4	—	15

---

*MN [1] = MSRI
[2] = MSEA

**PA has voluntary test at grade 11 (EQA) but at other grades has voluntary and mandatory. So coded mandatory information at other grade levels.

BEST COPY AVAILABLE

READING

LIST OF STATES FOR QUALITY INDICATORS PROJECT

				7 - 9						10 - 12						
		STATE		WA	VOC	LC	IC	SS	AT	WA	VOC	LC	IC	SS	AT	Sub-skill
		VERMONT			No program											
6a		VIRGINIA								"Reading - Min. Comp." - No other info)						
		WASHINGTON														
6a	20	WEST VIRGINIA	CTBS/U	5/1	40/2	2/1	43/13	20/3	—							
	20	WISCONSIN	CRT	—	—	?	?	?	—	—	--	?	?	?	--	18
			CTBS/U	5/1	40/2	2/1	43/13	20/3	—							
		WYOMING			No program											

BEST COPY AVAILABLE

APPENDIX 16

KEY: N = #s & Numeration
V = Variables
G = Geometry
M = Measure
S = Statistics

1 = Recall
2 = Manip.
3* = Explain (higher
4* = Prob. Slvg. order)

MATH

SUMMARY

KEY cont.: # of items/# of subskills
? = have but don't know
# of items

Source Rating	State		Gr. 1 - 3									4 - 6								
			N	V	G	M	S	1	2	3*	4*	N	V	G	M	S	1	2	3*	4*
4	ALABAMA	CRT	35/6	4/1	4/1	16/3	---	12/3	30/5	7/2	/0	52/10	3/1	14/2	15/4	4/1	21/4	41/9	17/4	9/1
		CAT	49/10	3/2	1/1	13/6	---	28/13	37/5	---	1/1	68/15	4/2	5/3	8/4	---	17/9	57/10	2/2	9/3
1	ALASKA		---									27/13	---	4/2	6/3	---	13/9	18/6	6/3	---
	ARIZONA	CAT	49/10	3/2	1/1	13/6	---	28/13	37/5	---	1/1	68/15	4/2	5/3	8/4	---	17/9	57/10	2/2	9/3
	ARKANSAS	CRT	52/13?	---	4/1	20/5	---	?	no information		---	52/13	---	8/2	12/3	---	?	no information		---
		SRA	---	---	---	---	---	---	---	---	---	57/10	---	4/3	9/3	---	18/9	43/4	5/1	4/2
3	CALIFORNIA		245/12	29/4	30/6	42/9	---	110/10	184/10	20/4	37/5	294/21	60/6	87/5	30/4	23/2	98/12	255/15	71/7	70/4
3	COLORADO																			
	CONNECTICUT	CAEP										39/6	4/1	3/1	14/3	---	21/5	33/4	6/2	---
		4th										272/17	48/3	18/1	64/5	---	144/9	152/10	88/6	16/1
		6th										100/10	12/3	8/2	20/4	---	40/8	56/10	28/6	16/4
	DELAWARE	CTBS/U	40/9	---	1/1	3/2	---	13/6	28/4	2/1	1/1	63/14	9/5	5/3	8/3	---	10/5	64/12	3/3	8/5
2	FLORIDA		49/5	---	---	19/3	---	20/4	48/4	---	---	88/8	5/1	---	23/3	---	29/4	83/7	4/1	---
5	GEORGIA		/8	/1	/1	/2	---	/8	/3	/1	---	/11	/3	/4	/2	/1	/5	/9	/7	---
1	HAWAII	SAT	85/11	6/1	5/4	9/3	---	29/9	58/5	5/2	13/3	96/13	6/1	5/2	10/1	1/1	118/18	51/7	9/3	21/3
		CRT			---- no info. ----															
	IDAHO																			
1	ILLINOIS											9/6	1/1	11/4	6/2	---	12/6	9/5	---	6/2
4	INDIANA		25/3	10/4	5/2	5/2	---	15/5	30/6	---	---	30/3	10/4	5/2	5/2	5/2	24/8	31/9	---	---
	IOWA																			
2	KANSAS		30/6	3/1	3/1	9/2	---	18/5	27/5	---	---	36/7	6/1	3/1	12/3	---	18/5	36/6	3/1 4th gr.	
												42/5	6/1	6/1	3/1	3/1	3/1	18/7	3/1 6th gr.	
	KENTUCKY	CTBS/U	40/9	---	1/1	3/2	---	13/6	28/4	2/1	1/1	63/14	9/5	5/3	8/3	---	10/5	64/12	3/3	8/5
2	LOUISIANA	2nd	52/8	---	---	4/1	---	36/7	16/1	4/1	---	60/6	8/2	12/2	8/2	---	24/6	56/5	8/1	---
		3rd	76/8	4/1	4/1	16/3	---	36/8	60/11	4/1	---									
	MAINE																			
5	MARYLAND	CAT	49/10	3/2	1/1	13/6	---	28/13	37/5	---	1/1	68/15	4/2	5/3	8/4	---	17/9	57/10	2/2	9/3
	MASSACHUSETTS																			
1	MICHIGAN											87/12	---	9/1	9/2	---	48/8	51/51	9/2	---
5	MINNESOTA	MEAM	133/8	---	16/2	31/2	---	?	can't tell		?	same (Minn. Basic Math)								
		BM	/15	/4	/1	/10	/0	/7	/16	/3	/4									

APPENDIX 16

BEST COPY AVAILABLE

MATH

Source Rating	State		Gr. 1 - 3									4 - 6								
			N	V	G	M	S	1	2	3*	4*	N	V	G	M	S	1	2	3*	4*
6	MISSISSIPPI		new - no info.									new - no info.								
2	MISSOURI											15/5	3/1	6/2	6/3	6/3	16/5	15/6	3/2	2/2
1	MONTANA											27/4	9/1	2/1	1/1	---	---	11/2	---	28/5
	NEBRASKA																			
	NEVADA	SAT	85/11	6/1	5/4	9/3	---	29/9	58/5	5/2	13/3	96/13	6/1	5/2	10/1	1/1	118/18	51/7	9/3	21/3
6	NEW HAMPSHIRE											/4	---	---	/2	---	/1	/3	/2	---
	NEW JERSEY																			
	NEW MEXICO	CTBS/U	40/9	---	1/1	3/2	---	13/6	28/4	2/1	1/1	63/14	9/5	5/3	8/3	---	10/5	64/12	3/3	8/5
3	NEW YORK		48/3	---	11/?	---	6/?	?	can't tell		?	44/3	---	13/?	---	9/?	?	can't tell		?
	NORTH CAROLINA	CAT	49/10	3/2	1/1	13/6	---	28/13	37/5	---	1/1	68/15	4/2	5/3	8/4	---	17/9	57/10	2/2	9/3
	NORTH DAKOTA																			
	OHIO																			
	OKLAHOMA																			
3	OREGON											67/11	---	---	2/1	---	2/1	37/4	13/4	17/3
5	PENNSYLVANIA											EQA-voluntary 36/14	2/1	7/3	12/4	1/1	23/9	30/11	2/1	3/2
		TELLS	/8	/1	/1	/4	/0	/6	/7	/0	/1	/6 TELLS	/1	/2	/3	---	/5	/6	---	/1
	RHODE ISLAND																			
6	SOUTH CAROLINA	CTBS/U	have M-CRT at grades 1,2,3,6 ---------------------									63/14	9/5	5/3	8/3	---	10/5	64/12	3/3	8/5
												---no other information ---------								
	SOUTH DAKOTA																			
	TENNESSEE																			
3	TEXAS		/7	---	/1	---	---	/4	/4	---	---	/5	/1	/1	/1	---	/2	/5	/1	---
	UTAH	CRT										63/14	9/5	5/3	8/3	---	10/5	64/12	3/3	8/5
		CTBS/S										80/15	5/3	5/1	13/3	---	14/7	68/10	6/2	15/3
	VERMONT																			
	VIRGINIA	SRA										57/10	---	4/3	9/3	---	18/9	43/4	5/1	4/2
	WASHINGTON	CAT										68/15	4/2	5/3	8/4	---	17/9	57/10	2/2	9/3
	WEST VIRGINIA	CTBS/U	40/9	---	1/1	3/2	---	13/6	28/4	2/1	1/1	63/14	9/5	5/3	8/3	---	10/5	64/12	3/3	8/5
	WISCONSIN	CTBS/U										63/14	9/5	5/3	8/3	---	10/5	64/12	3/3	8/5
	WYOMING																			

BEST COPY AVAILABLE

MATH

Source Rating	State		Gr. 7 - 9 N	V	G	M	S	1	2	3*	4*	10 - 12 N	V	G	M	S	1	2	3*	4*
4	ALABAMA	CRT	61/6	8/2	12/3	16/4	4/1	8/2	41/8	8/1	44/5	44/7	7/1	14/2	27/5	---	6/2	54/9	6/2	23/3
		CAT	66/17	7/6	9/4	7/3	---	11/6	60/14	2/2	16/8									
5	ALASKA		/10	/1	/1	/3	---		/5	/3										
	ARIZONA	CAT	66/17	7/6	9/4	7/3	---	11/6	60/14	2/2	16/8									
3	ARKANSAS	CRT	/26?	---	4/1	8/2	4/1	? no information ?												
3	CALIFORNIA		216/26	87/8	84/9	30/4	36/3	103/18	160/7	100/11	105/5	126/9	60/4	24/3	30/4	14/4	40/6	165/13	---	55/6
		other:	15/1																	
	COLORADO																			
	CONNECTICUT	CAEP	48/8	4/1	8/3	10/5	1/1	17/5	47/9	2/1	5/3	45/9	4/1	9/1	10/5	1/1	11/5	45/10	3/2	10/4
		Mastery	108/16	8/2	12/3	12/3	4/1	20/4	72/12	36/6	20/4									
		Prof.	47/13	3/2	4/3	10/3	1/1	8/4	43/12	6/3	4/3									
	DELAWARE	CTBS/U	61/15	13/7	4/3	7/3	---	10/6	64/16	1/1	10/5									
2	FLORIDA		95/6	4/1	---	15/2	---	---	84/7	20/1	10/1	35/2 80/7	5/1 ---	5/1 ---	25/5 ---	5/1 ---	5/1 20	30/4 60	--- ---	40/2 ---
5	GEORGIA		/11	/3	/2	/2	/1	/4	/10	/5	---	/10	/3	/3	/2	/1	/4	/10	/5	---
	HAWAII		SAT & DAT									HSTEC/3	---	---	/4 STAS	---	---	/5	---	/2
2	IDAHO		48/72	---	13/4	18/4	3/1	5/4	70/10	10/2	---									
1	ILLINOIS		12/7	4/3	26/12	7/2	---	6/4	24/16	2/3	17/3	6/4	7/2	23/8	5/2	---	7/4	15/5	3/1	16/5
4	INDIANA																			
	IOWA																			
2	KANSAS		48/8	---	3/1	3/1	3/1	9/2	39/7	---	9/2	42/1	3/1	6/2	6/2	3/1	---	9/3	---	51/3
	KENTUCKY																			
2	LOUISIANA		60/8	4/1	4/1	4/1	4/1	8/2	60/9	---	8/1	56/6	8/2	8/1	8/1	---	---	72/9	---	8/1
2	MAINE																			
5	MARYLAND	CRT	?	?	?	?	?	?	?	?	?									
		CAT	66/17	7/6	9/4	7/3	0	11/6	60/14	2/2	16/8									
	MASSACHUSETTS																			
1	MICHIGAN		81/11	3/1	12/3	12/2	---	39/8	60/7	9/2	---	72/9	3/1	15/3	12/8	6/2	12/3	78/11	12/3	6/2
5	MINNESOTA	BEAM	108/9	36/3	20/2	19/2	8/?	? can't tell ?			19/1?	91/10	51/4	30/2	20/2	8/?	? can't tell ?			29/3
		B.H.	/15	/4	/1	/10	/0	/7	/16	/3	/4									

BEST COPY AVAILABLE

MATH

Source Rating	State		Gr. 7 - 9									10 - 12								
			N	V	G	M	S	1	2	3*	4*	N	V	G	M	S	1	2	3*	4*
6	MISSISSIPPI		new - no info.									new - no info.								
2	MISSOURI											15/2	3/1	6/4	6/3	6/3	12/5	15/6	1/1	8/4
1	MONTANA											35/3	12/1	11/1	1/1	---		12/1	---	47/5
	NEBRASKA																			
2	NEVADA		30/5	12/1	6/1	46/5	---	10/3	74/6	4/1	6/2		same (9-12)							
6	NEW HAMPSHIRE		/6	/2	/2	/1	/3	/2	/10	/1	/1	/4	/2	/3	/1	/1	/2	/9	/0	/0
2	NEW JERSEY	(out)	65/9	5/1	12/2	10/6	---	11/3	57/8	5/2	19/5									
		(in)	57/10	9/2	11/3	15/2	1/1	5/2	48/10	9/2	31/4									
	NEW MEXICO	CTBS/U	61/15	13/7	4/3	7/3	---	10/6	64/16	1/1	10/5									
3	NEW YORK		/6	/2	/3	---	/1	/2	/7	/1	/2	same (9-12)								
5	NORTH CAROLINA	CRT										?	?	?	?	?	?	?	?	'?'
		CAT	66/17	7/6	9/4	7/3	---	11/6	60/14	2/2	16/8									
	NORTH DAKOTA																			
	OHIO																			
	OKLAHOMA																			
3	OREGON		59/12	---	---	---	---	---	34/5	10/4	15/3	58/9	---	---	---	---	2/1	35/5	1/1	20/2
5	PENNSYLVANIA	EQA-vol. IA	38/17	5/3	9/5	6/3	1/1	13/7	31/15	1/1	10/6	35/13	7/3	9/6	6/4	3/2	16/7	35/15	1/1	8/5
		TELLS	/12	/2	/4	/1	/2	/6	/12	/2	/1									
	RHODE ISLAND																			
6	SOUTH CAROLINA		have M-CRT at 8th - no info.									have M-CRT at gr. 11 - no info.								
		CTBS/U	61/15	13/7	4/3	7/3	---	10/6	64/16	1/1	10/5									
	SOUTH DAKOTA																			
5	TENNESSEE		/9	/1	/1	/3	/1	/3	/11	/1	---	same (9-12)								
3	TEXAS		/5	/1	/1	/1	/1	/1	/5	/0	/3									
	UTAH	CTBS/S										63/16	18/6	5/4	8/3	---	11/6	66/16	7/4	10/3
	VERMONT																			
	VIRGINIA											10th gr. math CRT - no other info.								
	WASHINGTON	CAT	66/17	7/6	9/4	7/3	---	11/6	60/14	2/2	16/8									
	WEST VIRGINIA	CTBS/U	61/15	13/7	4/3	7/3	---	10/6	64/16	1/1	10/5									
	WISCONSIN	CTBS/U	61/15	13/7	4/3	7/3	---	10/6	64/16	1/1	10/5									
	WYOMING																			

BEST COPY AVAILABLE

APPENDIX 17

WRITING

Source Rating	State		GRADES 1 - 3						GRADES 4 - 6					
			CO	GR	WU	OR	AT	SM	CO	GR	WU	OR	AT	SM
4	ALABAMA	CRT	42/5	--	13/3	4/1	--	--	43/5	--	20/4	16/3	--	--
		CAT	40/3	5/1	20/3	--	--	--	45/3	11/2	15/3	6/1	--	--
	ALASKA													
	ARIZONA	CAT	40/3	5/1	20/3	--	--	--	45/3	11/2	15/3	6/1	--	--
	ARKANSAS	SRA							54/3	5/2	25/5	--	--	--
3	CALIFORNIA		129/3	60/1	90/5	45/3	--	--	110/6	67/3	113/8	62/5	--	--
	COLORADO													
	CONNECTICUT	OLD							CAEP 5/3	--	12/4	3/2	19	8
		NEW							4th 21/4	--	15/3	--	--	1(H,A)
									6th /3	/1	/1	/2	--	1(H,A)
	DELAWARE	CTBS/U	2/1	8/2	10/2	--	--	--	70/5	14/3	17/4	10/2	--	--
3	FLORIDA		15/2	--	--	9/1	--	--	24/2	20/4	--	9/2	--	--
	GEORGIA													
	HAWAII	CRT		----- no info. ------										
		SAT	53/3	1/1	12/4	--	--	--	63/3	15/1	13/3	--	--	--
	IDAHO													
1	ILLINOIS								20/4	5/1	17/2	--	32	--
4	INDIANA(new)		?	?	--	--	--	1/1	?	?	--	--	--	?
	IOWA													
	KANSAS													
	KENTUCKY	CTBS/U	2/1	8/2	10/2	--	--	--	70/5	14/3	17/4	10/2	--	--
2	LOUISIANA		16/3	--	4/1	--	--	3/1 P						
	MAINE								--	--	--	--	--	X(H,P,A)
	MARYLAND	CAT	40/3	5/1	20/3	--	--	--	45/3	11/2	15/3	6/1	--	--
	MASSACHUSETTS													
	MICHIGAN													
	MINNESOTA													

KEY:

# of items/# of subskills
CO = conventions (e.g., spell, capit., punct.)
GR = Grammar (sentence structure)
WU = Word Usage
OR = Organization
AT = Attitude
SM = Writing Sample
? = Unknown # of items and subskills

BEST COPY AVAILABLE

WRITING

Source Rating	State	GRADES 1 - 3 CO	GR	WU	OR	AT	SM	GRADES 4 - 6 CO	GR	WU	OR	AT	SM
6	MISSISSIPPI	NEW						NEW					
2	MISSOURI							8/3	--	--	6/2	--	--
1	MONTANA							11/2	--	--	--	15	--
	NEBRASKA												
	NEVADA SAT	50/3	10/1	12/4	--	--	--	63/3	15/1	13/1	--	--	--
	NEW HAMPSHIRE												
	NEW JERSEY												
	NEW MEXICO CTBS/U	2/1	8/2	10/2	--	--	--	70/5	14/3	17/4	10/2	--	--
3	NEW YORK							--	--	--	--	--	2/(H)
	NORTH CAROLINA CAT	40/3	5/1	20/3	--	--	--	45/3	11/2	15/3	6/1	--	--
	NORTH DAKOTA												
	OHIO												
	OKLAHOMA												
3	OREGON							13/5	6/2	5/2	4/2	--	1/(H)
5	PENNSYLVANIA							5/2	20/3	5/2	7/3	--	--
	RHODE ISLAND												
	SOUTH CAROLINA CRT							W TESTED AT GRADE 6 - NO INFO					
	CTBS/U							70/5	14/3	17/4	10/2	--	--
	SOUTH DAKOTA												
5	TENNESSEE												
3	TEXAS	/4	/1	/1	--	--	1/1	/4	/1	/1	--	--	1/
	UTAH CTBS/S							70/3	--	24/5	11/2	--	--
	VERMONT												
	VIRGINIA SRA							54/3	5/2	25/5	--	--	--
	WASHINGTON CAT							45/3	11/2	15/3	6/1	--	--
	WEST VIRGINIA CTBS/U	2/1	8/2	10/2	--	--	--	70/5	14/3	17/4	10/2	--	--
	WISCONSIN CTBS/U							70/5	14/3	17/4	10/2	--	--
	WYOMING												

BEST COPY AVAILABLE

WRITING

Source Rating	State		GRADES 7 - 9						GRADES 10 - 12					
			CO	GR	WU	OR	AT	SM	CO	GR	WU	OR	AT	SM
4	ALABAMA	CRT	39/3	—	15/3	27/4	--	—	43/3	—	24/3	43/6	--	--
		CAT	45/3	14/3	12/3	11/4	--	--						
	ALASKA													
	ARIZONA	CAT	45/3	14/3	12/3	11/4	--	--						
	ARKANSAS													
3	CALIFORNIA		123/3	62/2	82/4	136/5	—	—	124/3	52/2		38/3	--	--
	COLORADO													
	CONNETICUT	Mstry	/3	/1	/1	/2	—	1(H,A)						
		Prof.	9/3	1/1	6/3	6/2	--	--						
		CAEP	29/3	6/3	40/6	17/7	41	11(?)	29/3	6/3	40/6	17/7	41	11(?)
	DELAWARE	CTBS/U	66/3	8/2	17/4	20/5	—	--						
3	FLORIDA		28/3	15/3	5/2	15/3	—	—	35/3	5/2	5/2	—	--	--
									--	--	--	20/2	--	--
	GEORGIA													
	HAWAII			SAT & DAT					CRT --	--	STAS --	/1	--	3(?)
2	IDAHO		21/1	—	--	--	—	(H)						
1	ILLINOIS*		2/1	5/1	13/1	14/1	32	—	2/1	5/1	13/1	14/1	32/	—
4	INDIANA (new)		?	?	—	--	--	?						
	IOWA													
	KANSAS													
	KENTUCKY	CTBS/U	66/3	8/2	17/4	20/5	--	--						
2	LOUISIANA		24/3	8/1	12/3	—	—	2(P)	40/3	12/1	4/1	4/1	—	2(P)
2	MAINE		—	—	—	—	—	?(H,P,A)	—	—	--	—	--	?(H,P,A)
5	MARYLAND	CRT	(NOT MUCH INFO)					2(?)						
		CAT	45/3	14/3	12/3	11/4	—	--						
	MASSACHUSETTS													
	MICHIGAN													
	MINNESOTA													
6	MISSISSIPPI		(NEW)							(NEW)				
2	MISSOURI								4/2	1/1	1/1	9/5	—	--
1	MONTANA								6/2	1/1	8/5	--	15/	--

---

*plus 16 "Mixed" items

BEST COPY AVAILABLE

WRITING

Source Rating	State	GRADES 7 - 9						GRADES 10 - 12					
		CO	GR	WU	OR	AT	SM	CO	GR	WU	OR	AT	SM
6	MISSISSIPPI	(NEW)							(NEW)				
	NEBRASKA												
2	NEVADA	—	—	—	—	—	2(H)		SAME (9 - 12)				
	NEW HAMPSHIRE												
	NEW JERSEY	12/3	18/3	12/4	24/4	—	3(H)		SAME (9 - 12)				
	NEW MEXICO CTBS/U	66/3	8/2	17/4	20/5	—	--						
3	NEW YORK	—	—	—	—	—	3(H)	—	—	—	—	—	3(H)
	NORTH CAROLINA CAT	45/3	14/3	12/3	11/4	—	--						
	NORTH DAKOTA												
	OHIO												
	OKLAHOMA												
3	OREGON	26/4	4/1	5/2	3/2	—	1(H)	—	—	—	—	—	2(H)
5	PENNSYLVANIA*	4/2	22/3	19/3	17/3	—	—	5/1	16/2	24/3	22/3	—	--
	RHODE ISLAND												
6	SOUTH CAROLINA CRT	W TESTED AT 8 - NO INFO						W TESTED AT Gr. 10 - NO INFO					
	CTBS/U	66/3	8/2	17/4	20/5	—	--						
	SOUTH DAKOTA												
5	TENNESSEE	/3	/3	/5	/1	—	--	SAME (9 - 12)					
3	TEXAS	/4	/1	/1	—	--	1						
	UTAH CTBS/S							50/3	—	25/5	10/2	—	--
	VERMONT												
	VIRGINIA SRA		NO INFO										
	WASHINGTON												
	WEST VIRGINIA CTBS/U	66/3	8/2	17/4	20/5	—	--						
	WISCONSIN CTBS/U	66/3	8/2	17/4	20/5	—	--						
	WYOMING												

---

*Voluntary

BEST COPY AVAILABLE

APPENDIX 18

SUMMARY OF NUMBERS OF ITEMS AND SUBSKILLS IN EACH CELL OF MATH MATRIX FOR GRADES 4-6 AND 4-9 IN CALIFORNIA, ALABAMA, FLORIDA, LOUISIANA, PENNSYLVANIA

KEY: Items/subskills

CALIFORNIA GR 6

MATH

	RECALL (facts, terms, symbols)	ROUTINE MANIP (compute simple word problems)	EXPLAIN (estimate, select algo, translate)	PROB SOLV (hard word probs, apply theorems)	TOTAL
Numbers	52/8	175/7	54/5	13/1	294/21
Variables	3/1	35/3	7/1	15/1	60/6
Geometry	43/3	12/1	0	32/1	87/5
Measurement	0	10/2	10/1	10/1	30/4
Statistics	0	23/2	0	0	23/2
TOTAL	98/12	255/15	71/7	70/4	494/38

CALIFORNIA GR 8

	RECALL	ROUTINE MANIP	EXPLAIN	PROB SOLV	TOTAL
Numbers	44/10	84/8	64/7	24/1	216/26
Variables	10/1	30/4	25/2	22/1	87/8
Geometry	45/6	6/1	7/1	26/1	84/9
Measurement	4/1	4/1	4/1	18/1	30/4
Statistics	0	36/3	0	0	36/3
Other				15/1 (prob solv w/ maps, signs, ads, schedules, charts)	15/1
TOTAL	103/18	160/17	100/11	105/5	468/51

BEST COPY AVAILABLE

MATH

ALABAMA GR 6

	RECALL	ROUTINE MANIP	JUDGE, TRANSLATE	PROB SOLV	TOTAL
Numbers	8/2	18/3	17/4	9/1	52/10
Variables	0	3/1	0	0	3/1
Geometry	9/1	5/1	0	0	14/2
Measurement	4/1	11/3	0	0	15/4
Statistics	0	4/1	0	0	4/1
TOTAL	21/4	41/9	17/4	9/1	88/18

ALABAMA GR 9

Numbers	0	21/3	8/1	32/2	61/6
Variables	0	4/1	0	4/1	8/2
Geometry	4/1	4/1	0	4/1	12/3
Measurement	4/1	8/2	0	4/1	16/4
Statistics	0	4/1	0	0	4/1
TOTAL	8/2	41/8	8/1	44/5	101/16

MATH

FLORIDA GR 5

	RECALL	ROUTINE MANIP	JUDGE, TRANSLATE	PROB SOLV	TOTAL
Numbers	25/3	59/4	4/1	0	88/8
Variables	0	5/1	0	0	5/1
Geometry	0	0	0	0	0
Measurement	4/1	19/2	0	0	23/3
Statistics	0	0	0	0	0
TOTAL	29/4	83/7	4/1	0	116/12

FLORIDA GR 8

Numbers	0	75/5	20/1	0	95/6
Variables	0	4/1	0	0	4/1
Geometry	0	0	0	0	0
Measurement	0	5/1	0	10/1	15/2
Statistics	0	0	0	0	0
TOTAL	0	84/7	20/1	10/1	114/9

MATH

LOUISIANA GR 4

	RECALL	ROUTINE MANIP	JUDGE, TRANSLATE	PROB SOLV	TOTAL
Numbers	12/3	40/2	8/1	0	60/6
Variables	4/1	4/1	0	0	8/2
Geometry	4/1	8/1	0	0	12/2
Measurement	4/1	4/1	0	0	8/2
Statistics	0	0	0	0	0
TOTAL	24/6	56/5	8/1	0	88/12

LOUISIANA GR 7

Numbers	8/2	44/5	0	8/1	60/8
Variables	0	4/1	0	0	4/1
Geometry	0	4/1	0	0	4/1
Measurement	0	4/1	0	0	4/1
Statistics	0	4/1	0	0	4/1
TOTAL	8/2	60/9	0	8/1	76/12

MATH

PENNSYLVANIA GR 5 "EQA" (voluntary in '84)

	RECALL	ROUTINE MANIP	EXPLAIN	PROB SOLV	TOTAL
Numbers	11/6	22/6	2/1	1/1	36/14
Variables	0	2/1	0	0	2/1
Geometry	4/2	3/1	0	0	7/3
Measurement	8/2	2/2	0	2/1	12/5
Statistics	0	1/1	0	0	1/1
TOTAL	23/10	30/11	2/1	3/1	58/24

PENNSYLVANIA GR 5 "TELLS" (number of items unspecified)

Numbers	/3	/3	0	0	/6
Variables	0	/1	0	0	/1
Geometry	/1	/1	0	0	/2
Measurement	/1	/1	0	/1	/3
Statistics	0	0	0	0	0
TOTAL	/5	/6	0	/1	/12

PENN GR 8 "EQA" (voluntary in '84)

Numbers	11/4	26/11	1/1	1/1	39/17
Variables	0	0	0	5/3	5/3
Geometry	5/2	3/2	0	1/1	9/5
Measurement	2/1	1/1	0	3/1	6/3
Statistics	0	1/1	0	0	1/1
TOTAL	18/7	31/5	1/1	10/6	60/29

MATH

PENNSYLVANIA GR 8 "TELLS" (number of items unspecified)

	RECALL	ROUTINE MANIP	EXPLAIN	PROB SOLV	TOTAL
Numbers	/2	/7	/2	/1	/12
Variables	/1	/1	0	0	/2
Geometry	/2	/2	0	0	/4
Measurement	/1	0	0	0	/1
Statistics	0	/2	0	0	/2
TOTAL	/6	/12	/2	/1	/21

KEY: Items/subskills

CTBS/U - Grade 6
DE, KS, NM,
SC, UT, WI

MATH

	RECALL	ROUTINE MANIP	EXPLAIN	PROB SOLV	TOTAL
Numbers	5/2	53/8	2/2	3/2	63/14
Variables	1/1	6/2	0	2/2	9/5
Geometry	4/2	1/1	0	0	5/3
Measurement	0	4/1	1/1	3/1	8/3
Statistics	0	0	0	0	0
TOTAL	10/5	64/12	3/3	8/5	85/25

APPENDIX 19
STQI Project Coding of Reporting Practices and Auxiliary Information

4/85

State:

Program: Minimum Competency Assessment (Testing)

Title of Document(s):

Description of Purpose of Report:

- Audience:
- Authoring Agency:
- Authors:
- Date of Report:
- Stated Purpose or Objectives:
- Type of Report: Results Technical

I. General Description

A) Type of Test: Commercial (i.e., Published Standardized Measure from Vendor)

Private (i.e., Privately or Internally Developed)

B) Name of Test:

C) Version or Edition:

D) Enter Dates of Testing (Month and Periodicity) and Nature of Testing (Census, Sample)

Subject Area	Grade Level K	1	2	3	4	5	6	7	8	9	10	11	12
Reading													
Math													
Language Arts													
Writing													
Other:													

II. Reported Results

A) Metric:

1. Indicate the type of scale(s) used to report results:

______ Raw Score
______ Scale Score (define:____)
______ Percent Correct
______ Grade Equivalent
______ Percentile Rank
______ NCE's
______ Stanine
______ Other

2. Which of the above are most frequently used? ______________________________

____

3. Enter Statistic (Measure of Central Tendency) used for Reporting

Grade Level

Metric	K	1	2	3	4	5	6	7	8	9	10	11	12
Raw Score													
Percent Correct													
Grade Equiv.													
Percentile Rank													
Scale Scores:													
Stanines													
NCE's													
Z-Scores													
T-Scores													
Other:													

B) Student Subgroup Definitions (Check all that apply)

- Racial/Ethnic Groups (List groups identified in Report)

  ____

  ____

  ____

- Sex

- Special Programs (List all progress identified in Report)

  ____

  ____

  ____

- Language Status (List all groups identified in Report)

  ____

  ____

- Other (List characterisitics and groups identified in Report)

  ____

  ____

C) School/District Groupings (Check all that apply and enter groups identified)

		School	District
____	Size	____________	____________
____	Geographic Location	____________	____________
____	Program Types	____________	____________
____	Socio-Economic	____________	____________
____	Other (specify)	____________	____________
	____________	____________	____________
	____________	____________	____________

D) Descriptive Statistics

(Enter grade levels and subject areas at which the descriptive statistic cited is given for each grouping.)

	All Students	Type 1 subgroups ( )	Type 2 subgroups ( )	Type 3 subgroups ( )	Type 4 subgroups ( )
I. Central Tendency Measures					
Mean					
Median					
Mode					
Other(Name)					
________					
II. Variability/ Dispersion Measures					
Stand.Dev.					
Variance					
Range					
Other(Name)					
________					
III. Distributional Information					
Quartile					
Duciles					
Quintiles					
IV. Frequencies of students attaining each:					
Raw Score					
% Correct					
Other					
V. Percentages of students attaining each:					
Raw Score					
% Correct					
Other					
Cut. Point: above					
below					

E) Longitudinal Information:

- Longitudinal Information Present   Yes   No
- Cohort Reported:
  - Same Students (Other________) tracked____________________
  - Same Grades (Different Students) ____________________
  - Other:
- Period for Reported Data: ____________________
- Number of Time Points: ____________________
- Periodicity (How often conducted): ____________________
- Type of Statistic Reported at each Point:
  - Measure of Central Tendency: ____________________
    ____
  - Measure of Variability: ____________________
    ____
  - Other: ____________________
    ____
- Test/Measure Stability:
  - Number of years with same measure: ____________________
  - Name of current measure: ____________________
  - Name of previous measure (if any): ____________________
  - Nature and reason for change: ____________________
    ____

F) Supplemental Analyses:

- Psychometric Analyses: 1) Item Analyses: Difficulty
  - Point ?
  - Item Characteristic Information
  - Other:
- 2) Reliability: Internal Consistency
  - Split-Half
  - Test-Retest

3) Validity: Concurrent

Content

Predictive

Construct

- Factor Analyses: (describe use) ________________________________

  ____

- Curricular Match: (describe) ________________________________

  ____

- Test Bias Analyses: (describe) ________________________________

  ____

- Teacher Analyses: (describe) ________________________________

  ____

G) Volume of Data:

- Number of Students Tested: Total______

  Per Grade Level_______

  ____

  ____

H) Non-Test Information Collected

1. Types of Additional Information Collected:
   - Attitudinal: ________________________________
   - Demographic: ________________________________
   - Other: ________________________________

     ____
2. Level (Respondent Level) of Information Collected:
   - Student
   - School
   - District
   - Other

I) Other Reports Available from Progress:

J) Other Comments:

APPENDIX 20

ADDENDUM

Linking State Educational Assessment Results: A Feasibility Trial

Prepared by R. Darrell Bock
National Opinion Research Center, University of Chicago
November, 1985

Recent developments in the technology of educational measurement present opportunities for obtaining comparative information on educational progress in the states. This concepts paper reviews some of these advances and outlines a proposed feasibility trial of one of them.

1. Background

Although the sample surveys conducted by the National Assessment of Educational Progress (NAEP) provide accurate measures of educational outcomes for the nation as a whole, the sampling rates are too low to enable reporting for geographical areas smaller than the four main regions -- Northeast, Southeast, Midwest and West. As a result, no between-state comparisons of outcomes, or comparisons of state results with the national average, are possible within the present budgetary limitations of NAEP. Several strategies exist, however, for obtaining such information. One that has already been proposed is for states to bear the cost of extending the NAEP sample to enough students from their schools to insure a dependable state average. As a very rough estimate, the marginal cost to each state for the additional sampling might be $150,000.

States that already have system-wide attainment testing programs in operation could, however, obtain comparable or better information at less cost by making use of item-response theoretic (IRT) methods for linking of test scales (Lord, 1980). These methods would permit the states to express the scores of their present tests on a common scale, which could be linked to the NAEP scale. The equating procedures require only that a small number of common, or "anchor" items, from each of the state tests be present in a specially prepared equating test that is administered to a broadly representative group of students at the relevant grade level. The scaling of items in this equating test can then be propagated back to the state test in order to define a scale with common origin and unit of measurement in all of the test. If scaled NAEP items are also included, the common scale can be related to NAEP results. Apart from this one-time study establishing the equating links (which would need to be repeated when a state's test changed), the annual scoring of state results on the common scale would be a straightforward computer operation costing perhaps $100 per 10,000 students.

There are two possible approaches to creating the special equating tests:

1.1 For those states that are already testing closely similar subject matter, the simplest approach is for them to contribute to the equating test three to six of their items in each skill area for which scales are to be constructed. These items, plus some scaled NAEP items in the same areas, would then be administered under uniform conditions in a few selected schools in the participating states. Since the results would be used only in test linking and not for describing attainment, the sampled schools would not need to be representative of the state. It is only necessary that the full range of student attainment is covered. the data obtained in this way in the participating states would then be collated for IRT scaling. Similar scaling of the state test from which these items arose would also be carried out separately on operational data supplied by each state. The item scale parameters of the anchor items would then be used to adjust all of the state results to the same origin and unit of measurement. Using these results, each state could express the attainment of pupils or schools in therms of this common scale. All participating states' results would then be comparable and could be related to the corresponding NAEP scale if NAEP items were included. Even commercial tests could be included in the linking, provided the publishers would agree to this use of some of their items.

1.2. If the states are not already testing in comparable subject-matter skill areas, a more extensive initial effort would be required. Curriculum experts from each of the participating states would have to meet and agree on the content of the areas to be tested. They would then have to assemble and select items representing this content. Some new items might have to be written, but for the most part existing items from state testing programs and from NAEP could be used. This newly constructed equating test would then be administered to a broad sample of students at the relevant grade level and the results subjected to IRT analysis as above. Each state could then insert some of the scaled items from the equating test into new tests devised for its own program, by scoring the new tests by IRT methods anchored on these scaled items, each state could then express its outcome measures on the same scale for purposes of comparison with other states or with national results.

In addition to the economy of these linking strategies for comparing educational outcomes in the states, they have several advantages offer the alternative of extending the NAEP sample: 1) no additional operational testing beyond that of the existing state program would be required, 2) the state would have results for all students included in the existing state program, not just those in the probability sample collected by NAEP, 3) the objectives and content of the state testing would not be determined or limited by NAEP policy and practices in assessment, 4) commercial as well as state testing organizations could participate, 5) new avenues for communication between the state testing programs would be opened, and the capabilities of the programs would be strengthened, and 6) in the course of choosing content and skills to be included in the equating, greater consensus between the states on curriculum problems would be fostered.

2. Proposal for an Initial Feasibility Trial

Results of recent study by Burstein, et al, (1985) reveal sufficient communality of test content at the eighth grade level to support a trial of the first of these two linking methods in a number of states. It is proposed that five of these states join in a pilot study to evaluate procedures for this purpose and to develop prototypes of documents for reporting and comparing state educational outcomes. The study would be limited to measures of 1) reading proficiency and 2) basic mathematical skills, assessed in three schools in each of these states during the spring term. A high, middle and low SES school should be enlisted for this purpose by each of the respective state education offices. Each school would be requested to make one fifty minute class period available for administration of the equating test to all or most of their eighth grade students.

The states should be selected to include at least one that employs traditional individual student achievement testing and one that employs matrix sampled assessment. In addition at least one of the states should routinely test in the autumn in grade eight and one in the spring of grade seven. States on both plans present a special problem in equating because the scores from the earlier testing or different grade level must be adjusted to their predicted values for the standard testing time and grade level tested. So that corrections of scores for nontypical testing time can be estimated, those states not testing in the spring of grade eight should then test all students in the pilot schools in both grade eight and grade seven.

Each state would contribute four items each from its current reading and mathematics tests for grade eight. NAEP would be requested to provide an additional four scaled items in each of these subject-matter areas. These items would be assembled into a 48-item expendable-form test intended for non-speeded adiministration.

Coordination of the testing and monitoring of test administration in each school would be handled by field staff of a national survey organization.

Scoring and IRT analysis of the resulting data would be contracted to an organization with capabilities in this area. Each state would also supply this organization with a computer tape containing the response of students to items of its reading and mathematics tests administered in current operational testing. The latter data would be IRT scored on the common scale for purposes of the prototype demonstration of between-state comparisons and relating to the NAEP national results. The organization or organizations responsible for field testing and analysis would produce the prototype report and also submit a technical report documenting procedures and discussing any significant problems encountered during their work.

Because of its experimental nature, this proposed trial has been held to modest proportions to keep costs low. It is estimated that, once the states agree to cooperate and the items for the equating test have been assembled, the field work and analysis could be carried out by an organization already equipped for these activities for about $80,000 of direct costs.

3. Further Steps

Procedures for the proposed initial trial are sufficiently straightforward that a three month lead time should be enough to prepare the test and make arrangements for field testing. Another three months should be enough for analysis and preparation of the prototype report. If the feasibility trial is judged successful, work could begin on an operational system involving more subject-matter areas. At that point, it is likely that the participating states will wish to move to the second strategy for linking based on the development of a common equating test. Some of the states might then choose to alter their testing programs to conform more closely to the content of that test. Such changes, supported by the scale linking through the equating test, would further facilitate the comparison of educational outcomes among the states and with the nation as a whole.

REFERENCES

Burstein, L., Baker, E.L., Aschbacher, P. & Keesling, W. (1985). Using Test Data for National Indicators of Educational Quality: A Feasibility Study. Los Angeles: Center for the Study of Evaluation, UCLA Graduate School of Education.

Lord, F.M. (1980). Applicatons of Item Response Theory to Practical Testing Problems. Hilsdale, NJ: Earlbaum.